Multiword Units in Russian Everyday Speech: Empirical Classification and Corpus-Based Studies

Natalia V. Bogdanova-Beglarian[1[0000-0002-7652-0358]], Olga V. Blinova[1,2[0000-0002-5665-3495]], Maria V. Khokhlova[1[0000-0001-9085-0284]], Tatiana Y. Sherstinova[2,1[0000-0002-9085-3378]], and Tatiana I. Popova[1[0000-0003-2066-7868]]

[1] Saint Petersburg State University, Saint Petersburg, Russia
[2] HSE University, Saint Petersburg, Russia
{n.bogdanova, o.blinova, m.khokhlova, t.sherstinova, t.i.popova@spbu.ru}@spbu.ru

Abstract. The article is dedicated to the results of a research project describing the classes and functioning of multiword units in contemporary Russian everyday speech. The concept of multiword units encompasses quite diverse linguistic phenomena, making the creation of a working typology one of the project's central tasks. This typology is necessary for annotating corpus material and obtaining statistical characteristics. The identified classes of multiword units include the following units: 1) non-phraseologized collocations, 2) phraseologized collocations, 3) occasional collocations, 4) idiom forms, 5) constructions, 6) precedent texts and their elements, 7) multi-word pragmatic markers, and 8) speech formulas. The article describes the methods for annotating these units using the ORD corpus of everyday spoken Russian and presents the results of a quantitative analysis of their functioning within the annotated subcorpus. The obtained data can be used to address both theoretical tasks in the field of lexical and grammatical description of Russian everyday speech and numerous tasks related to processing or generating live spoken Russian.

Keywords: modern Russian, everyday speech, oral discourse, multiword units, collocations, syntax, statistical analysis, speech corpus, corpus linguistics, speech technologies.

1 Introduction

The study of spoken language, especially using a corpus approach based on recordings obtained in natural communication settings, reveals phenomena that are not reflected in existing dictionaries and grammars but actively function in the speech of native speakers. These phenomena, therefore, require special documentation and analysis for linguistic and language-teaching purposes, as well as for creating human-like dialogue systems and artificial intelligence. This research focuses on phenomena at the intersection of vocabulary, grammar, and syntax, which we refer to collectively as *multiword units*. A number of works are dedicated to their analysis and study, for example, [1]; [2]; [3]; [4]; [5]; [6]; [7]; [8]; [9]; [10]. These units require not only

theoretical description but also the formation of an inventory—a lexicon (in the broadest sense of the word), in which they would be represented with the necessary quantitative characteristics. For the codified Russian written language, multiword units are relatively fully described. However, in the class of multiword units used in spontaneous spoken speech, there are still many "white spots" despite the emerging linguistic and digital resources. For example, databases exist for the Russian language that combine units of different types: multiword units, collocations, constructions, etc. ([11]; [12]; [13]; [14]; [15]; [16]; [17], etc.).

The relevance of addressing the problem outlined is connected to the fact that, recently, linguistics has become closely intertwined with information technology and the development of various speech applications. Solving these tasks requires not only a coherent theory but also a large volume of annotated linguistic data. The study of multiword units and their identification in a speech corpus involves addressing issues related to lemmatization and the representation of their morphological, syntactic, and semantic features.

The material for studying multiword units is the ORD corpus of Russian everyday speech, characterized by the fact that recordings are obtained in natural communicative situations [18]; [19] and reflect the full richness and diversity of everyday speech communication—in terms of the topics of conversation [20], participants [21], and communication conditions [22].

The approaches to searching for and identifying multiword units are diverse, relying on both expert and automatic techniques. In our study, it seems reasonable to employ both. The initial list of multiword units was obtained through expert methods, followed by n-gram analysis of transcriptions of everyday spoken language to identify the most frequent bigrams and trigrams. The results of the n-gram analysis were described in [23]; [24]. However, the overall number of n-grams obtained, amounting to tens of thousands of units, and the lack of context pose an obvious drawback for subsequent expert work. Therefore, in this study, the basis for collecting multiword units was their expert manual annotation, described below in Section 3, and relying on their empirical classification presented in Section 2. The article also presents the results of automatic clustering of the empirically obtained list of units (see Section 4). Finally, Section 5 provides preliminary statistics on the distribution of multiword units based on the study sample.

2 Empiric Classification of Multiword Units

The concept of *multiword units* encompasses a wide range of linguistic phenomena, so the creation of their working typology is one of the central tasks of the project. This typology is essential for the subsequent annotation and processing of the material. The typology of multiword units was developed in several iterations. In the first stage, a pilot classification of multiword units was used, focusing on their structural and lexical features. Based on the results of comprehensive pilot annotation of oral speech transcriptions, taking into account the main proposed types of multiword units,

this typology was revised, and a new scheme was proposed. This scheme currently includes eight main categories:

1. Non-phraseologized collocations,
2. Phraseologized collocations,
3. Occasional collocations,
4. Idiom forms,
5. Constructions,
6. Precedent texts and their elements,
7. Multi-word pragmatic markers,
8. Speech formulas.

Non-phraseologized collocations are stable combinations whose perception does not determine the imagery of the meaning.

Phraseologized collocations are stable constructs whose elements possess figurative meanings. As a result of the interaction between the semantics of the construction components, a certain meaning is fixed in spoken language for the unit: "ne obrashchat' vnimaniya" ("to ignore"), "v poryadke veshchey" ("as a matter of course"), "doyti do ruchki" ("to reach the limit"). This type of multiword unit is closest to traditional phraseological units.

Occasional collocations, as the name suggests, are modifications of commonly accepted collocations in the language.

An **idiom form** is considered a word form that, due to frequent use, acquires functional and semantic significance in everyday communication (most often this is a prepositional-case form of nouns): for example, "po ponyatiyam" ("according to the rules"), "do figa" ("a lot"), "ne v kaif" ("not enjoyable"), "v printsipe" ("in principle"), "ne po sebe" ("uncomfortable"), "ne gorit" ("not urgent"), and others.

The concept of a **construction** differs from an idiom form and a phraseologized collocation in that the structure of the construction includes a constant component and a variable component X: <X ni razu ne Y> ("X never Y"), <X-u ne do Y-a> ("X doesn't care about Y"), <nu + Acc!> ("come on + Acc!"), etc.

Elements of precedent texts refer to fragments of well-known phrases, for example, from movies: "ikh yest' u menya" ("I have them") (a phrase from Lev Slavin's play "Intervention"), "chey tuflya" ("whose shoe") (from Leonid Gayday film "Kidnapping, Caucasian Style"), etc.

Pragmatic markers are functional units of oral discourse that help speakers structure dialogue and mark speech intention. Pragmatic markers often have a complex structure, thereby expanding the list of multiword units: "ya ne znayu" ("I don't know"), "skazhem tak" ("let's say"), "kak govorit'sya" ("as they say"), "tak skazat'" ("so to speak"), "nu vot" ("well then"), "i vse dela" ("and all that"), "ili kak eto" ("or whatever"), etc.

Speech formulas are often interjectional units that reflect the speaker's emotional reaction or a response in a dialogue: "vot yeshchyo!" ("there you go! "), "nichego sebe!" ("wow!"), "kak khochesh'" ("as you wish"), "kak znayesh'" ("as you know").

The results of the pilot annotation showed that the proposed classification generally well reflects the features of multiword units characteristic of spoken language,

therefore it is accepted as the main one for conducting expert annotation of these units and further research.

3 Expert Annotation of Multiword Units in ORD corpus

The ORD corpus is a complex and multi-component resource used for conducting research on Russian spoken discourse at all linguistic levels. An important result of the ongoing research is the manual expert annotation of the corpus materials at the level of multiword units.

3.1 Multiword Units Annotation Principles

The annotation of multiword units in the ORD corpus is carried out as follows. Experts review the transcriptions of audio recordings, which are exported into a tabular format, and fill in the multiword units database using a form that includes the following fields:

1. Communicative episode,
2. Speaker code,
3. Phrase,
4. Multiword unit as it appears in the text,
5. Class of the multiword unit according to the proposed typology (the *Tags* column in the database),
6. Invariant (optional — filled in only when the form of the initial multiword unit was clear without doubt),
7. New multiword unit — a note indicating whether the multiword unit was included in the initial list.

The following annotation codes were proposed:

1. Non-phraseologized collocations — NK,
2. Phraseologized collocations — FK,
3. Occasional (non-conventional) collocations — OK,
4. Idiom forms — ID,
5. Constructions — KS,
6. Elements of precedent texts <text> — PT,
7. Multi-word pragmatic markers — PM,
8. Speech formulas — RF.

In the *Tags* column, it was preferable to record only one (the main) variant of the multiword unit's characteristic, as, unlike pragmatic markers of spoken language, the units under study are not prone to multifunctionality; they are primarily annotated in terms of their formal organization.

3.2 Multiword Units Annotation Results

Four experts participated in the annotation process, one of whom (E1) acted as the curator and made final corrections. The episodes of natural speech communication selected for annotation varied in duration, and thus, differed in the volume of text transcriptions. Therefore, multiple speech episodes were selected for some informants, affecting the distribution of material among the experts for annotation. Each episode was annotated by one expert. The final distribution of annotated episodes was as follows:

E1 – 14 episodes (7.8%);
E2 – 20 episodes (10.26%);
E3 – 101 episodes (51.79%);
E4 – 60 episodes (30.77%).

The experts thoroughly reviewed the text transcriptions in the *Phrase* column of the research database and recorded information about the multiword units found in the phrases in columns created specifically for annotation. The multiword units were recorded in the form they appeared in the fragment in a specially designated column (*Multiword units*).

In total, 195 macro-episodes were annotated, with a total volume of 300,000 word usages. The manual annotation of multiword units enabled the creation of an expanded list of these units, preliminary statistical information on the implementation of these units in spoken language, and the identification of the main difficulties in expert annotation of these units.

3.3 Main Challenges in the Annotation Process of Multiword Units

The main difficulties in the annotation process arose with those multiword units that were not included in the initial list, necessitating collective decisions on the classification of each unit and whether the unit could be considered a multiword unit at all.

A particularly close connection was found between constructions and non-phraseologized collocations, as the lack of imagery and figurative meaning of the components distanced the unit under study from phraseologized collocations. In such cases, the determining factor was the search for a variable (X) that is defining from the conceptual framework's perspective.

Another problem in the annotation of multiword units was that some word combinations primarily realized their grammatical meaning and syntactic valence rather than stability and lexicalization, which prevented them from being classified as multiword units, even though their combination could potentially be considered regular.

At the final stage of annotation, it became clear that identifying the invariant for each realization of multiword units is a separate research task. For example, in constructions, it was necessary to determine the fixed part and its form, and then the part that is variable. Next, the grammatical and lexical characteristics of the potential variable needed to be described.

In speech formulas, some components may be perceived as optional, but upon analysis, it becomes clear that only the full composition of the multiword unit realizes

its meaning. For example, the unit "da ladno" ("oh, come on") can serve as a reaction precisely in this structural variant because identifying "da" ("oh") as an optional part "(da) ladno" ("come on") causes the unit to cease being a multiword unit and loses its function of expressing the speaker's reaction.

Some units annotated as phraseologized collocations also exhibit variability. For example, the multiword units "morochit' golovu" ("to mess with someone's head") and "vynosyt' mozgi" ("to blow someone's mind") can be perceived as synonymous, or their proximity can be seen as a potential to fill positions with words of a certain meaning, allowing the multiword units to be considered as constructions. Only further expert work and linguistic analysis will allow the formation of a final list of invariants for each realization of multiword units and the creation of a new classification of multiword units in terms of their formal organization.

4 Automatic Clustering of Multiword Units

The empirically derived list of multiword units was subsequently subjected to an automatic clustering procedure. Initially, automatic clustering of multiword units was carried out based on the results of expert annotation of oral speech transcriptions for a sample of 300,000 tokens. The clustering was performed using the k-means algorithm without considering metadata but utilizing two different approaches to data vectorization: 1) tf-idf (CountVectorizer from sklearn) and 2) FastText embeddings. Calculations were performed for models with 5, 10, 15, and 30 clusters[1]. During automatic clustering, the elements within each cluster were grouped around one or more keyword features.

The most semantically meaningful clusters were obtained when the sample was divided into 30 clusters. For example:

CLUSTER #8 multiword units:
[*'million raz' ('a million times'), 'desyat' raz' ('ten times'), 'inoy raz' ('sometimes'), 'pervyy raz slyshu' ('first time I hear it'), 'sto raz' ('a hundred times'), 'paru raz' ('a couple of times'), 'kak raz' ('just right'), 'lishniy raz' ('one more time')*]

CLUSTER #22 multiword units:
[*'ne moyo' ('not my thing'), 'ne pozhalela deneg' ('didn't spare the money'), 'ne problema' ('no problem'), 'ryadom ne stoyat' ('don't come close'), 'darom ne nuzhna' ('don't need it for free'), 'ne bum-bum' ('don't get it'), 'ne svetit' ('not gonna happen'), 'ne sud'ba' ('not meant to be'), 'ne govorite' ('don't say'), 'sovest' ne gryizla' ('didn't feel guilty'), 'nikak ne doberus'' ('can't get around to it')*]

[1] The choice of the maximum value of the number of clusters depends on the volume of the analysed data. In the first experiment the results of manual annotation of multiword units were clustered, while in the second experiment the lists of n-grams obtained automatically were processed, hence, the volume of data in the second case was larger.

Nevertheless, automatic clusters can sometimes contain an "exception". For example, cluster #19 (when dividing multiword units into 30 clusters) mainly consists of units containing the lemma "delo" ("thing" or "matter"). However, for some reason, the borrowed English multiword unit "vi a ze chempions" ("we are the champions"), which belongs to the type of precedent texts, also ended up in this cluster.

CLUSTER #19 multiword units:
[*'odno delo' ('one thing'), 'sovsem drugoe delo' ('a completely different matter'), 'obychnoe delo' ('a usual thing'), 'imeyu delo' ('have a matter'), 'khoroshee delo' ('a good thing'), 'delo khoroshee' ('the matter is good'), 'takie dela' ('such things'), 'ponyatnoe delo' ('obviously'), 'poslednee delo' ('the last thing'), 'delo poshlo' ('the matter progressed'), 'strannoe delo' ('a strange thing'), 'takoe delo' ('such a thing'), 'sereznoe delo' ('a serious matter'), 'svyatoe delo' ('a sacred thing'), 'vi a ze chempions' ('we are the champions'), 'temnye dela' ('dark matters'), 'drugoe delo' ('another matter'), 'bylo delo' ('there was a matter')*]

At the next stage of data processing, automatic clustering was performed for the complete list of frequent n-grams, where *n* takes a value from 2 to 5 for the entire volume of existing oral speech transcripts of the ORD corpus. The clustering was conducted using the k-means algorithm without considering metadata, but utilizing two different approaches to data vectorization: tf-idf and FastText embeddings. Given the large number of units studied, amounting to tens of thousands of unique types for each of the 2-, 3-, 4-, and 5-grams, it was decided to divide the research sample into 50 clusters.

For each n-gram size, four files were obtained — two txt files with clusters (where key features are highlighted as a separate line for the tf-idf model) and two csv files with complete lists of types and the cluster number in the second column. These tables allow for the analysis of statistics and, if necessary, enable the data to be traced back to the original sources.

The results show that the n-gram clusters differ significantly from each other in structural and semantic cohesion. See, for example, clusters 1 and 45 for bigrams:

CLUSTER #1

Types: [*'ya priedu' ('I will come'), 'ya zabyl' ('I forgot'), 'ya rabotayu' ('I am working'), 'ya poprobuyu' ('I will try'), 'ya chitayu' ('I am reading'), 'ya reshil' ('I decided'), 'ya yezdila' ('I went'), 'ya vspomnila' ('I remembered'), 'ya rad' ('I am glad'), 'ya kupila' ('I bought'), 'ya skhozhu' ('I will go'), 'ya ya' ('I I'), 'ya poprosil' ('I asked'), 'kak ya' ('like I'), 'ya vozmu' ('I will take'), 'ya yeye' ('I her'), 'ya polozhila' ('I put'), 'naskol'ko ya' ('as far as I'), 'ya napishu' ('I will write'), 'kotoroye ya' ('which I')*]

CLUSTER #45

Types: [*'ugu kogda' ('mm-hmm when'), 'mam a' ('mom uh'), 'nado tuda' ('need to go there'), 'on yey' ('he her'), 'zdes' bylo' ('here was'), 'ya dolzhen' ('I must'), 'podozhdi a' ('wait uh'), 'vidite kak' ('see how'), 'togda davay' ('then let's'), 'tri shtuki' ('three pieces'), 'interesno ya' ('interesting I'), 'dumala ya' ('thought I'), 'moemu a' ('my uh'), 'ta m'*

('that um'), 'a yey' ('uh her'), 'ya im' ('I them'), 'ugu u' ('mm-hmm uh'), 'e potomu' ('uh because'), 'ponimayesh' ty' ('you see'), 'vy tozhe' ('you too')]

Even greater diversity is observed for larger n-grams. The conclusion that can be drawn from this study is that, in the future, clustering should be performed not on the entire array of n-grams obtained, but only on the most frequent units (the upper zone of the n-gram frequency dictionary).

The study showed that automatic clustering, with a correctly selected number of classes, is a useful tool for the preliminary grouping of word sequences based on their lexicon. Since automatic clustering relies solely on the lexical composition of multiword units without considering semantics, expert analysis is necessary for further work with such data. A useful property of a cluster is its reliance on "key" word(s), which allows grouping similar multiword units and can be used to search for invariant forms. For example:

CLUSTER #30 (when dividing multiword units into 30 clusters)

multiword units: [*'vot eto vot' ('this one here'), 'vot ona vot' ('here she is'), 'vot eti vot' ('these ones here'), 'vot beda' ('what a trouble'), 'vot etot vot' ('this one here'), 'vot tebe' ('here you go'), 'vot eti samye' ('these very ones'), 'vot eta bol' vot' ('this pain here'), 'vot takiye vot dela' ('that's how things are'), 'vot takoy vot' ('this kind of'), 'vot etu vot' ('this one here'), 'vot takiye vot' ('these kinds of'), 'vot takiye dela' ('these are the things'), 'vot tuda vot' ('over there'), 'vot takaya vot' ('this kind of'), 'vot imenno' ('exactly'), 'vot etim vot' ('with these here'), 'vot tak vot' ('that's how it is')*]

It can be assumed that this property of clusters will be most pronounced with a sufficiently large number of them. However, this hypothesis requires experimental verification.

Regarding the clustering of a large number of automatically obtained n-grams, the analysis showed that the results do not have a distinguishing function that would be useful for the automatic identification of multiword units, at least for the counting methodology used in the project. This problem might be resolved by machine learning methods based on expert selection, but for this task, the volume of expert annotation needs to be significantly expanded.

5 Preliminary Statistics of Multiword Units Distribution in Everyday Conversations

In the course of the study, statistical data were obtained on the conditions of the realization of multiword units in everyday spoken language and their distribution in specific types of communicative macro-episodes, as well as in relation to other communication conditions. A description of the obtained statistics was also provided.

5.1 Most Frequent Multiword Units

The overall frequency of use of multiword units in a representative sample was obtained. The total lexicon of multiword units identified from the ORD material during manual annotation (on a subsample of 300,000 words, 195 speech episodes) amounted to 1,088 units of various types (see section 2).

The results showed that the composition of these most frequent stable multiword units in our everyday communication is quite heterogeneous.

The most frequent unit "V PRINTSIPE" ("in principle") (rank 1) is a lexicalized prepositional-case form (idiom form) or a pragmatic marker (verbal hesitative or delimiter, primarily navigational, depending on the context).

Similarly, the unit "V OBSHCHEM" ("generally") (rank 10) in this frequency list can be characterized. The idiom form "V OBSHCHEM" ("generally") as a pragmatic marker is a verbal hesitative, delimiter of all three types (initial, navigational, and final), and occasionally a self-correction marker, also depending on the context.

From the class of pragmatic markers in the top 10, there are also units "ETO SAMOE" ("you know") (rank 2) (verbal hesitative, self-correction marker, delimiter marker of all three types (initial, navigational, and final), and rarely a xenopointer marker), "NA SAMOM DELE" ("actually") (rank 5) (verbal hesitative), and "I TAK DALEE" ("and so on") (rank 7) (placeholder marker).

Thus, 50% (exactly half) of the most frequent multiword units in our spoken communication are primarily pragmatic markers, which are not included in this status in traditional explanatory dictionaries, including dictionaries of Russian colloquial speech, nor in the "Russkiy konstruktikon" [25], nor in the "Pragmatikon" [26]. All data on these markers (their functional characteristics) are provided here according to the Dictionary of Pragmatic Markers [17]. Two of the 5 units of this type ("V PRINTSIPE" ("in principle") and "V OBSHCHEM" ("generally")) are also idiom forms, constituting a separate class of multiword units. This polyfunctionality is characteristic of many spoken language units, reflecting the overall diffuse nature of this material.

The remaining units that made it into the top 10 are speech formulas (40%) ("NICHEGO SEBE" ("wow"), "SLAVA BOGU" ("thank God"), "DA TY CHO" ("really"), "VSE RAVNO" ("anyway")), included in the "Pragmatikon" since they are predominantly response replicas in dialogue, and a phraseologized collocation (10%) ("VSE VREMYA" ("all the time")), definitely included in the "Russkiy konstruktikon".

It should also be noted that all the most frequent multiword units in our everyday speech are bi- and trigrams, described in [23]; [24].

5.2 Most Frequent Classes of Multiword Units

The top 5 of this frequency list include phraseologized collocations (rank 1), idiom forms (rank 2), speech formulas (rank 3), pragmatic markers (rank 4), and syntactic constructions (rank 5) ("delo v tom chto" ("the fact is that"), "v lyubom sluchaye" ("in any case"), etc.).

The analysis showed that among the phraseologized collocations, the most commonly used units in everyday Russian speech are "VSE VREMYA" ("all the time") (3.67%) and "PONYATNOE DELO" ("obviously") (2.20%) (percentage calculated within each group); the most frequent idiom form is "V PRINTSIPE" ("in principle") (34.55%).

For speech formulas, the top 4 ranks include the same units "NICHEGO SEBE" ("wow"), "SLAVA BOGU" ("thank God"), "DA TY CHO" ("really"), and "VSE RAVNO" ("anyway"), which are in the top 10 of the overall frequency list of multiword units.

For the group of pragmatic markers (PM), again, the top 4 positions are occupied by units from the overall frequency list of multiword units ("ETO SAMOE" ("you know"), "NA SAMOM DELE" ("actually"), "I TAK DALEE" ("and so on"), "V OBSHCHEM" ("generally")). It is also evident that the obtained data reflect the frequency of realizations of multiword units, not their base variants (invariants). In the lexicon of Russian Pragmatic Markers [17], the realizations "ETO SAMOE" and "ETOT SAMYI" are one marker "ETO SAMOE" ("you know") (this "classic" form is the most frequent in our speech and is used in any hesitative search, including when grammatical adjustment to the desired noun is not required); the realizations "VOT TAK VOT" and "VOT ETO VOT" are also one deictic marker "VOT (...) VOT" ("this one here"), which exists exclusively as a structural model that is filled each time with a new unit: "VOT TAK VOT" ("this way"), "VOT TAKOY VOT" ("this kind of"), "VOT OTSYUDA VOT" ("from here"), etc. This marker simply does not have a single base (standard) form, which is why it occupies a special place in the lexicon of pragmatic markers. Neither dictionaries nor grammars of the Russian language highlight this construction as an independent unit, whereas corpus material analysis shows its very high frequency (rank 19 in the list of 60 Russian pragmatic markers).

Among syntactic multiword unit constructions, the most common are "V LYUBOM SLUCHAYE" ("in any case") and "DELO V TOM CHTO" ("the fact is that") (4.88% each), among non-phraseologized collocations are "ODNU SEKUNDOCHKU" ("one moment") (5.38%), as well as "DRUGOE DELO" ("another matter"), "PO KRAYNEY MERE" ("at least"), and "CHEGO-TO TAKOE" ("something like that") (4.62% each). Again, it is clear that this refers only to specific realizations of multiword units. For example, alongside "ODNU SEKUNDOCHKU" the lexicon contains "ODNU SEKUNDU" ("one second") (rank 56). However, the expected invariant form "CHTO-TO TAKOE" ("something like that") was not found next to "CHEGO-TO TAKOE". This once again indicates that the question of multiword unit invariants is not as simple as it seems at first glance and requires separate consideration.

Multiword units from the classes of occasional collocations and precedent texts are predictably rare. Interestingly, a significant portion of occasional collocations units include obscene vocabulary, although such vocabulary is also present in other groups. Overall, both of these classes of multiword units provide good material for analysis from various perspectives.

5.3 Part-of-Speech Composition of Multiword Units

The entire material of the annotated subcorpus (a subsample of 300,000 words from 195 speech episodes) was automatically tagged for the part-of-speech (POS) of the components of multiword units, allowing for the generation of frequency lists based on this parameter.

The most frequent POS structure turned out to be PREP NOUN (a noun with a preposition, lexicalized prepositional-case word form, or idiom form) (14.85%). The most typical units of this type are: "V PRINTSIPE" ("in principle") (40.59%), "V SMYSLE" ("I mean") (5.61%), "V ITOGE" ("as a result") (3.96%), "PO IDEE" ("supposedly") (3.63%).

Other frequent structures are ADJF NOUN (a combination of a full adjective (including adjective-pronoun and numeral-pronoun) with a noun) (5.98%) and PREP ADJF NOUN (the same combination with a preposition) (5.78%). The most typical units of these two types are: "PONYATNOE DELO" ("obviously") (9.84%), "ODNU SEKUNDOCHKU" ("one moment") (5.74%), "DRUGOE DELO" ("another matter") and "KAKAYA RAZNITSA" ("what's the difference") (4.92% each); "NA SAMOM DELE" ("actually") (23.73%), "V LYUBOM SLUCHAE" ("in any case") (11.86%), "DO SIKH POR" ("up to now") (6.78%), "VO VSYAKOM SLUCHAE" ("anyway") and "PO KRAYNEY MERE" ("at least") (5.08% each).

5.4 Frequency of Use of Multiword Units Depending on Speakers' Social Characteristics

The research sample included speech episodes from 111 informants' speech days, among which there were 57 women and 64 men. The sample also included the speech of their 727 interlocutors, among which there were 645 women and 272 men. More than 50% of the material studied involved domestic communication, with business communication being the second most frequent.

These data correlate with information about the social roles of the speakers: most often, speakers took on the role of "friend", with the second most common role being "work colleague".

It has already been noted that multiword units from occasional collocations and precedent texts classes are predictably rare. Interestingly, occasional collocations multiword units (33 instances in the material) are used equally by both women and men: 17 uses in women's speech and 16 in men's speech. Of the 23 instances of precedent texts, 14 are used by women and only 9 by men.

The use of multiword units from the Non-phraseologized collocations, constructions, and pragmatic markers classes is relatively evenly distributed among women (60% of uses) and men (40% of uses). The use of phraseologized collocations is slightly more common among women (55%), while speech formulas are more characteristic of women's speech (68%).

The distribution of multiword units across age groups does not have striking features, as the percentage distribution is relatively even. Only a few indicators stand out:

- Older men use idiom form multiword units less than middle and younger age groups;
- In the speech of older women, speech formulas are predominant.

The level of speech competence is determined in the ORD corpus through the correlation of two indicators: the level of education and the professional activity of the informant. The results indicate that the use of various classes of multiword units is generally more characteristic of people with an intermediate level of speech competence (only 5 to 20% among people with a high level of speech competence).

Other features of the use of multiword units in different communication situations were also identified and described.

6 Conclusion

The study presents a typology of multiword units for spontaneous everyday Russian speech and provides statistical data on their realization based on the manually annotated subcorpus of the well-known ORD corpus. Due to the labor-intensive nature of manual annotation, only one-third of the existing transcriptions in the corpus have been annotated to date. Therefore, the presented statistics should be considered preliminary, and the study of multiword units continues along the following paths: 1) by expanding the volume of annotated data to 1 million word usages and 2) by involving automatic analysis tools for processing multiword units [27].

Methods for automatically identifying stable multiword units will rely on existing lexicons, but due to the homonymy of linguistic units, they will require subsequent manual correction. Special scripts are being created to search for new forms of constructions [28] based on invariant structures of multiword units. In addition, modern speech technologies allow for a significant expansion of the empirical base of corpus research by attracting new representative volumes of audio recordings. Such work is currently being carried out on the materials of the ORD corpus [29]; [30], and conducting statistical analysis of multiword units on extended volumes of transcriptions will allow for the correction of quantitative data on their usage in different communication situations by different types of speakers.

The obtained data can be used to address both theoretical tasks in the field of lexical and grammatical description of Russian everyday speech and numerous tasks related to processing or generating live spoken Russian. Additionally, the research results will form the basis of a Dictionary of Collocations and other multiword units of everyday Russian speech.

Acknowledgments. This research has been carried out thanks to the financial support of Russian Science Foundation (project No. 22-18-00189 "Structure and Functionality of Stable Multiword Units in Russian Everyday Speech").

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Columbus, G.: Processing MWUs: Are different types of MWUs psycholinguistically valid? An eye-tracking study. In David Wood (ed.), Perspectives on formulaic language incommunication and acquisition, 194–210. New York: Continuum (2010)
2. Moon, R.: Vocabulary Connections: Multi-word Items in English. In N. Schmitt & M. McCarthy, (Eds.), Vocabulary: Description, Acquisition and Pedagogy, pp. 40–63. Cambridge: Cambridge University Press. (1997)
3. Moon, R.: Frequencies and forms of phrasal lexemes in English. In Anthony P.Cowie (ed.),Phraseology: Theory, analysis, and applications,79–100. Oxford: ClarendonPress. (1998)
4. Nattinger, J., DeCarrico, J.: Lexical phrases and language teaching. Oxford: Oxford University Press (1992)
5. Nunberg, G., Sag, I., Wasow, Th.: Idioms.Language70(3), 491–538 (1994)
6. Schweigert, W.: The comprehension of familiar and less familiar idioms. Journal of Psycholinguistic Research15, 33–45 (1986)
7. Weinreich, U.: Problems in the analysis of idioms. In Jaan Puhvel (ed.),Substance and structure of language,23–81. Berkley: University of California Press (1969)
8. Wray, A.: Formulaic language and the lexicon. Cambridge: Cambridge University Press. (2002)
9. Wray, A.: Formulaic language: Pushing the boundaries. Oxford: Oxford UniversityPress. (2008)
10. 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: Digital Geography. Proceedings of the International Conference on Internet and Modern Society (IMS 2022). Springer, Part F2317, pp. 129–139 (2024)
11. Bast, R., Endresen, A., Janda, L. A., Lund, M., Lyashevskaya, O., McDonald, J., Mordashova, D., Nesset, T., Rakhilina, E., Tyers, F. M., Zhukova, V.: The Russian Constructicon. An electronic database of the Russian grammatical constructions. (2021) https://constructicon.github.io/russian/ last accessed 2024/7/15
12. Janda, L. A., Lyashevskaya, O., Nesset, T., Rakhilina, E., Tyers, F. M.: Chapter 6. A Constructicon for Russian: Filling in the Gaps. In Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, & Tiago Timponi Torrent (eds.), *Constructicography: Constructicon development across languages* [Constructional Approaches to Language 22], 165–181. Amsterdam: John Benjamins Publishing Co. (2018) DOI: https://doi.org/10.1075/cal.22.06jan
13. Khokhlova, M.: Collocations in Russian Lexicography and Russian Collocations Database. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 3198–3206. Marseille, France. European Language Resources Association (2020)
14. 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), pp. 43–45. Sofia: INCOMA Ltd (2015).
15. Khokhlova, M. Attributive collocations in the gold standard of Russian collocability and their representation in dictionaries and corpora. Voprosy Leksikografii, 21, 33–68. (2021)
16. 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).
17. Pragmatic markers of Russian everyday speech: Dictionary-monograph / Ed. N.V. Bogdanova-Beglaryan. St. Petersburg: Nestor-History (2021).
18. Asinovsky, A., Bogdanova, N., Rusakova, M., Stepanova, S., Ryko, A., Sherstinova, S.: The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // Lecture Notes in Computer Science – Vol. Text, Speech and Dialogue, – № 5729/2009. (2009).
19. 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. (2009).
20. Akinshina, E., Sherstinova, T.: Thematic Diversity of Everyday Russian Discourse: a Case Study based on the ORD corpus. In: Mahadeva Prasanna et al. (eds), Specom 2022, LNCS 13721, Springer Nature, 2022. Pp. 1-9. (2022).
21. 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. (2016).
22. 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, pp. 268–276. (2015).
23. Khokhlova, M., Blinova, O., Bogdanova-Beglarian, N., Sherstinova, T.: On the most frequent sequences of words in Russian spoken everyday language (bigrams and trigrams): an experience of classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). SPECOM 2023, 14338 LNAI, pp. 455–466 (2023)
24. Sherstinova, T., Markovich, O.: N-gram Analysis of Everyday Russian Speech: in Search of Multiword Units. In: 35th Conference of Open Innovations Association (FRUCT), 2024, April 24-26, Tampere, Finland. Pp. 831-838. (2024).
25. Russkiy konstruktikon: https://constructicon.github.io/russian/
26. Pragmatikon: https://pragmaticon.ruscorpora.ru
27. Sherstinova, T., Popova, T.: Multiword Units in Russian Spoken Language: Methods for Lexicon Expansion and Statistical Analysis. LiLaC (under review) (2024).
28. Rakhilina, E. V.: Lingvistika konstrukciy (Construction Linguistics) / Ed. E. V. Rakhilina. Moscow: Azbukovnik Publishing Center. (2010).
29. Sherstinova, T., Kolobov, R., Mikhaylovskiy, N.: Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level. In: Proceedings of SPECOM 2023 / LNCS, 14338/14339, pp. 43–56.
30. Sherstinova, T., Mikhaylovskiy, N., Kolpashchikova, E., Kruglikova, V.: Bridging Gaps in Russian Language Processing: AI and Everyday Conversations. In: 35th Conference of Open Innovations Association (FRUCT), 2024, April 24-26, Tampere, Finland. Pp. 253-258.