Modeling Imperative Utterances in Russian Spoken Dialogue: Verb-Central Quantitative Approach

Olga Blinova(✉)

Saint Petersburg State University,
7/9 Universitetskaya nab., St. Petersburg 199034, Russia
o.blinova@spbu.ru

Abstract. The study is aimed at detecting stable wording patterns of the utterances with directive function in Russian, and based on the material of speech corpus containing long-term audio recordings of everyday spoken communication. The lemmatized and morphologically annotated mini-corpus in question includes 2030 utterances with 2nd person Sg and Pl verb forms in imperative mood and consists of 11075 word forms. The research involves the data on frequencies of (co-)occurrences of word forms, lemmas, parts of speech within the mini-corpus.

Keywords: Speech corpus · Russian · Everyday speech · Spoken dialogue · Imperative utterances · Pragmatics · Dialogue acts

1 Introduction

This paper presents some results of the description of imperative utterances in the corpus of Russian everyday communication «One day of speech» (ORD). By now the ORD includes transcripts and multi-level linguistic annotation for audio recordings representing daily speech by 127 informants and their numerous interlocutors. Creation principles of the corpus are described in detail in [1].

The main goal of the ORD corpus creation is to fix Russian spontaneous speech in natural communicative situations, and to get authentic data from everyday speech and spontaneous interaction. Face-to-face dialogues are the main part of the corpus. The linguistic material of such type is especially well suited for the studies in real linguistic behavior, in particular, for the analysis of the ways of "doing things with words" in Austin's sense [2].

2 Illocutionary Force Indicating Devices or Dialogue Act Cues

One of the central challenges of the corpus pragmatics is the dialogue acts annotation. «Dialogue act» (DA) loosely means a «speech act used in dialogue» [3]. The important question is: what kind of search approaches can be used in a corpus for identification of given linguistic expressions as utterances with certain pragmatic meaning?

There are the features for marking the illocutionary force of utterances, the so called «illocutionary force indicating devices», IFIDs. According to [4] IFID is «any element of a natural language which can be literally used to indicate that an utterance of a sentence containing that element has a certain illocutionary force or range of illocutionary forces». We can use standard IFIDs as search strings. There are lexical, morphosyntactic and prosodic IFIDs, such as lexemes, word order, intonation [5]. However, there are speech acts, which ostensibly «do not appear in routinized forms or in reliable combination with IFIDs», and directive speech acts in English fall into this category [6].

As reported by D. Jurafsky, there are lexical, syntactic, prosodic, and discourse cues for dialogue act identification, including in particular lexical cues, so called 'cue phrases' [7]. The inventory of cross-linguistically common lexical or syntactic cues for imperatives (commands) includes particles, verbal clitics, special verb morphology, subject omitting etc. [7, 8]. In Russian we can distinguish several types of devices which serve to indicate the type of dialogue act, first of all, it is a grammatical mood.

3 Types of Form/Function Correspondence in Russian Verb Utterances

As A. Aikhenvald explains, «imperative mood is the commonest way of expressing commands in languages of the world» [9]. In terms of theory of speech acts, commands belong to the group of directives [10]. Using directive, the speaker tries to cause the hearer to do or not to do something. Russian has a special morphological imperative forms, and the imperative occurs in its prototypical directive function above all.

Imperative forms also can occur in 'transposed' uses, «which are not directive in the prototypical sense but only express directivity in a very weakened form» [11]. E. Fortuin in [12] speaks about

(1) necessitive use: *Все ушли, а я сиди дома* [13, §1948]
all gone but I-Nom sit-Imp.2Sg at home
'Everybody has gone, but I have to stay at home'

(2) narrative use: *Мы возьми и напиши на сайт президенту* [14]
we-Nom take-Imp.2Sg and write-Imp.2Sg to the website to the president
'We suddenly wrote to the website to the president'

(3) optative use: *Награди вас господь за вашу добродетель* [12, 162]
reward-Imp.2Sg you-Acc god-Nom for your goodness
'May God reward you for your goodness'

(4) conditional use: *Будь я помоложе, и позволила бы комплекция, сам бы полез* [12, 177]
be-Imp.2Sg I-Nom younger and allowed Irr bodily constitution, self Irr climb
'Had I been younger, and had my bodily constitution allowed it, I would have climbed myself'

(5) concessive use: *В какую сторону ни гляди, выхода нет* [12, 216]

in which side not look-Imp.2Sg escape not
‘No matter in which direction you look, there is no way out’.

There are semantic-syntactic features, which can provide identification of directive versus non-directive uses of imperative. Some relevant features are: aspect, possibility of expressing subject, occurrence of the suffix *-те*, presence of particle *-ка*, presence of particle *бы*, word order [12]. Thus, we can involve the information about co-occurrences of imperative forms for the semantic qualification of the imperative utterances.

In Russian, **commands also can be regularly expressed by the non-imperative verb forms**, by using: (1) present tense forms *В эти игры ты больше не играешь* ‘Don’t play these games any more’, (2) future tense forms *Пойдёшь со мной* ‘Go with me’, (3) past tense forms *Пошёл отсюда* ‘Get off’, (4) infinitive *Стоять* ‘Stay put’, ‘Freeze!’, (5) irrealis *Сходил бы ты в магазин* ‘Maybe you should go to the shop’.[1]

N. Stojnova in the paper devoted to imperative uses of indicative present and future forms in Russian [16] indicates, that there are some formal features, which can mark pragmatic similarity of the non-imperative utterance to prototypical directive use of the imperative. So, there are certain patterns of non-imperative commands formation. The features she mentions are as follows: presence of the subjective pronouns *ты*, *вы* etc., occurrence with particles of the type *ну-ка*, aspect.

Thus, the three possibilities of form/function correspondence for the verb utterances have been identified: directive imperatives, non-directive imperatives, and non-imperative directives. Certain morphological, morpho-syntactic and lexical features can indicate pragmatic meaning of the utterance created on the basis of imperative or non-imperative verb form.

4 Study Design, Material and Method

The actual study is aimed at detecting stable wording patterns of the utterances with directive function, and detection of formal markers, which can indicate pragmatic meaning of a directive.

The subcorpus used for this research encompasses mainly face-to-face dialogues between 42 informants and their interlocutors which include 240000 word forms. The paper concentrates on utterances in the imperative mood with the verb in the second person Sg or Pl. All utterances of this kind were extracted from a subcorpus. The lemmatized and morphologically annotated mini-corpus includes 2030 imperative utterances and consists of 11075 word forms. So, the mini-corpus here under analysis is composed of the imperative utterances only. The mini-corpus in question is small, but highly homogeneous: it consists of the utterances with prototypical imperative forms mainly in prototypical directive function.

As D. Jurafsky indicates, the simplest way to build a probabilistic model for detection of lexical and phrasal cues (resp. lexical and morphosyntactic IFIDs)

[1] For details, see [15]. In the listing the so called «whimperatives» and other indirect ways of expressing commands, as well as verbless directives are not taken into account.

«is simply to look at which words and phrases occur more often in one dialogue act than another» [7, 597]. N-gram model is used successfully in practical implementations of dialogue act detection, e.g. yes-no-questions in English often have bigram sequences of the type *do you*, *are you*, *was he* (or trigram sequences of the type <*start*> *do you*) [7, 17].[2]

The actual research is based on the information about occurrences of word forms, lemmas, parts of speech, inflectional forms, and about co-occurrences of word forms, parts of speech, inflectional forms within the exploratory subcorpus of directives. Thus, frequency-ordered lists of the unigrams, lemmas, POSs, some forms of inflection, as well as lists of most common bigrams with an imperative component are considered.

An utterance in the actual research usually is a fragment of the text transcript between two marks of phrasal division '//', '?' et al. However, in the mini-corpus there are many single-word utterances of the type *слушай*, *послушай* 'listen', *смотри* 'look', *подожди* 'wait'. E.g., the following phrase is divided by two parts – the part consisting of the attention getting device, the imperative *слушай*, and subsequent statement: *Слушай в буфете я не беру сосиски* // 'Listen at the lunchroom I don't take the sausages'.

5 Results and Discussion

Firstly, the data on frequency distribution of POS classes was obtained.[3] The most frequent parts of speech in the mini-corpus are: the verb, the particle, the noun, the pronoun. It is worth noting the high position of the particle in the list (Table 1).

Secondly, the list of most commonly used colligations (including colligations with verbs in the imperative mood in the second person Sg or Pl) was created. The list is based on bigram co-occurrence data of tags of POS classes. Table 2 lists top ten colligations. The data demonstrate in particular a high degree of co-occurrence between imperative forms and particles (which use in front of the verb), another imperative forms, and nouns (which usually use in front of the verb).

Thirdly, the lists of most frequent unigrams and most frequent lemmas were created. The stopword list at present includes prepositions only, and does not include particles, conjunctions, pronouns etc. The stop words are *в*, *на, у, с, к, по* etc.

The list of top thirty most frequent unigrams includes: PART *ну* 'well' (#332), NEG PART *не* (#278), PART *вот* (#269), imperative which usually functions as an attention-getting device *слушай* 'listen-Imp.2Sg'(#228), SPRO *ты* 'you-Nom.Sg'

[2] See [18] for a detailed overview of the approaches to dialogue act recognition, based on intra-utterance features or on inter-utterance context.

[3] Morphological annotation is carried out using the analyzer MyStem, developed for Russian by I. Segalovich and V. Titov at «Yandex». The list of POS-tags includes: S = noun, A = adjective, NUM = numeral, ANUM = numeral adjective, V = verb, ADV = adverb, PRAEDIC = predicative, SPRO = pronoun, APRO = adjectival pronoun, ADVPRO = adverbial pronoun, PR = preposition, CONJ = conjunction, PART = particle, INTJ = interjection, COM = part of compound word; «foreign» means a word of a foreign language. The abbreviations NEG = negative (negation), IRR = irrealis are used in the glosses above.

Table 1. Frequency distribution of POS classes

POS	Count	Percent	POS	Count	Percent
V	3142	28,37	APRO	308	2,78
PART	1752	15,82	A	239	2,16
S	1574	14,21	INTJ	209	1,89
SPRO	1418	12,80	NUM	120	1,08
PR	650	5,87	ANUM	45	0,40
ADV	641	5,79	Foreign	2	0,02
CONJ	603	5,45	COM	2	0,02
ADVPRO	370	3,34	**Total**	**11075**	**100**

Table 2. Commonly used colligations

Colligation	Count	Illustration	Translation
PART + V-Imp-2	589	*ну смотри*	Well look
V-Imp-2 + V-Imp-2	443	*иди иди*	Go go
S + V-Imp-2	408	*Маш бери*	Masha-Voc take
V-Imp-2 + PART	405	*слушай ну*	Listen well
PART + PART	328	*ну вот*	Well
V-Imp-2 + SPRO	313	*подожди ты*	Wait you-Sg
V-Imp-2 + S	311	*дай ложку*	Give a spoon
PR + S	300	*в холодильник*	In the fridge
S + PART	268	*Коля ну*	Kolya well
SPRO + V	266	*я говорю*	I say

(#200), CONJ *и* 'and' (#195), SPRO *я* 'I-Nom' (#185), CONJ, INTJ or PART *a* (#164), PART or CONJ *да* (#148), SPRO or CONJ *что* (#144), PART, APRO or CONJ *так* (#130), SPRO, PART or APRO *это* (#124), SPRO or APRO *все* (#115), PART or ADVPRO *там* (#114), V *смотри* 'look-Imp.2Sg' (#105), SPRO *мне*, 'I-Dat', 'to me' (#103), V *подожди* 'wait-Imp.2Sg' (#103), PART or V *давай* 'let's', 'give-Imp.2Sg' (#95), INTJ *э* (#94), SPRO *вы* 'you-Nom.Pl' (#94), PART *пожалуйста* 'please' (#87), ADVPRO, PART or CONJ *как* (#70), ADV *сейчас* 'now'(#69), SPRO *меня* 'I-Gen (Acc)' (#57), PART, SPRO or CONJ *то* (#56), V *иди* 'go- Imp.2Sg'(#55), V *смотрите* 'look-Imp.2Pl' (#49), PART *нет* (#47), PART or ADV *еще* (#44), *он* 'he-Nom' (#44).

The data obtained show a predominant use of a range of particles including *ну*, *вот*, *так*, *давай*. The presence of most common polite formula *please* should be noted and the absence of the post verbal particle -*ka* in the list of top 30 unigrams. The subject pronoun *я* 'I' appears in the occurrences of the type *я и говорю*, *я тебе говорю*, *я же говорю* 'I'm saying', *я сказала* 'I said' et al., see the following example: *не лезь к девочке/я тебе/десять раз уже сказала* // 'Do not bother the girl/I told you ten times already'.

As it was expected, the frequency list of verb forms shows a predominance of imperative. However, among frequently occurring verb forms there are two indicative forms: the one mentioned above *говорю* ‘I’m saying’, and *хочешь*, which usually occurs in the form, as in the utterance *хочешь ночуй/хочешь уезжай//* ‘You can stay if you want or leave if you want’.

The most frequent lemmas that have inflected forms are represented in Table 3 below.

Table 3. Top-thirty most frequent lemmas

Rank	Lemma	Count	Rank	Lemma	Count
1	*я* ‘I’	337	15	*идти* ‘go’	64
2	*ты* ‘you-Sg’	267	16	*извинить* ‘excuse’	62
3	*слушать* ‘listen’	252	17	*мы* ‘we’	57
4	*смотреть* ‘look’	155	17	*давай* ‘let’s’	57
5	*вы* ‘you-Pl’	151	18	*они* ‘they’	50
6	*это* ‘it’	133	18	*говорить* ‘say’	50
7	*быть* ‘be’	127	19	*давать* ‘give’	49
8	*подождать* ‘wait’	118	20	*взять* ‘take’	37
9	*он* ‘he’	106	21	*делать* ‘do’	36
10	*все* ‘everyone’	103	22	*такой* ‘such’	33
11	*сказать* ‘say’	99	23	*знать* ‘know’	30
12	*она* ‘she’	82	23	*держать* ‘hold’	30
13	*посмотреть* ‘look’	67	23	*хотеть* ‘want’	30
14	*дать* ‘give’	65	24	*мочь* ‘can’	29
14	*этот* ‘this’	65	25	*написать* ‘write’	28

The lemma *я* ‘I’ has the leading position due to the large amount of entries of the type *дай мне* ‘give it to me’, *позвони мне* ‘call me’, *скажи мне* ‘tell me’, *послушай меня* ‘listen to me’ etc. Lemmas *ты*, *вы* represent subjective pronouns (those that serve as non-omitted subjects) above all.

Fourthly, the list of bigram sequences on word forms was created. Table 4 lists 20 most frequent bigrams with verbs in the second person Sg or Pl in the data. As it can be seen, sequences with particles predominate, while sequences with content words encompass only insignificant part of the list.[4]

Whereas the most frequent bigrams may be considered as collocation candidates, the values of t-score for the most frequent two-word sequences were counted. Some sequences with relatively high t-score are not fully compositional, indeed. Thus, *говорить* with negation (t-score 3,99) is used in the utterances of the type *и не говори* ‘don’t even say this’ that usually express agreement with the other communicant: *ага// вот именно//не говори* ‘yes//sure//don’t even say this’.

[4] The sequences *слушай слушай* ‘listen listen’(#148), *подожди подожди* ‘wait wait’(#33) and *слушайте слушайте* ‘listen-Pl listen-Pl’(#17) are not under consideration, as their presence in the data is caused by the multiplicity of single-word utterances *слушай*, *слушайте*, *подожди*.

Table 4. Most frequent bigram sequences on word forms

2-gram	Count	2-gram	Count
не говори ‘don’t say’	16	*ну слушай* ‘well listen-Sg’	6
ну смотри ‘well look-Sg’	15	*ну смотрите* ‘well look-Pl’	6
вот смотри ‘here look-Sg’	15	*дай ей* ‘give-Sg it to her’	6
сейчас подожди ‘now, wait-Sg (just a minute)’	11	*иди иди* ‘go-Sg go-Sg’	6
слушай ну ‘listen-Sg well’	11	*извини меня* ‘excuse-Sg me’	6
вот смотрите ‘here look-Pl’	9	*не лезь*‘don’t meddle’	5
иди сюда ‘come-Sg here’	9	*ну посмотри* ‘well look-Sg’	5
скажите пожалуйста ‘tell-Pl please’	9	*ну попробуй* ‘well try-Sg’	5
скажите а ‘tell-Pl me’	8	*ну расскажи* ‘well tell-Sg’	5
ну подожди ‘well wait-Sg’	7	*вы посмотрите* ‘have-Pl a look at it’	5

6 Conclusion

The results of the study confirm the significant role of «small words» in wording of the utterances with imperatives in directive function. Thus, most frequent parts of speech in the mini-corpus of directives are the verb and the particle; the most frequent unigrams are the particles *ну*, *не*, *вот*; the sequences with particles predominate in the list of most frequent bigram sequences. By now it is clear that the features, which can indicate pragmatic meaning of a directive in Russian, are the colligations of the type *ну* + V-Imp-2, *вот* + V-Imp-2, *пожалуйста* + V-Imp-2.

Imperative forms in ‘transposed’ uses can hardly demonstrate such sequential patterns. Thus, the incorporation of *пожалуйста* ‘please’ in all types of ‘transposed’ uses looks equally unacceptable, cf.: **Все ушли а я дома пожалуйста сиди*, **Мы пожалуйста возьми и напиши* <…> etc.[5] Combinations with most common particles *ну* and *вот* ‘well’ seem to be acceptable to varying degrees, cf.: **Все ушли, а я ну сиди дома*, but *Все ушли а я вот сиди дома*. However, these findings need to be verified with the use of corpus material. It is also worth noting that we do not know, whether the ‘transposed’ uses occur in the colloquial speech, or in the fiction texts and academic grammars only. Consequently, the directions of further study are: the improvement of using n-gram model due to addition of the position numbering, and the corpus study of ‘transposed’ uses of imperative forms.

Acknowledgement. The research is supported by the Russian Science Foundation (RSF), project #14-18-02070 «Everyday Russian Language in Different Social Groups».

[5] The asterisk marks unacceptable sentences.

References

1. 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. LNCS, vol. 5729, pp. 250–257. Springer, Heidelberg (2009)
2. Austin, J.L.: How to Do Things with Words, 2nd edn. Oxford University Press, Oxford (1976)
3. Bunt, H.: The dit++ taxonomy for functional dialogue markup. In: Decker, S., Sichman J., Sierra, C., Castelfranchi, C. (eds.) Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), pp. 13–20 (2009)
4. Searle, J.R., Vanderveken, D.: Speech acts and illocutionary logic. In: Vanderveken, D. (ed.) Logic, Thought and Action, pp. 109–132. Springer, Heidelberg (2005)
5. Weisser, M.: Landmarks in ‘Traditional’ Pragmatics. http://martinweisser.org/courses/intro/pragmatics.html
6. Flöck, I., Geluykens, R.: Speech acts in corpus pragmatics a quantitative contrastive study of directives in spontaneous and elicited discourse. In: Romero-Trillo, J. (ed.) Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies, pp. 7–37. Springer, Heidelberg (2015)
7. Jurafsky, D.: Pragmatics and computational linguistics. In: Horn, L., Ward, G. (eds.) The Handbook of Pragmatics, pp. 578–604. Blackwell Publishing, Oxford (2006)
8. Sadock, J.M., Zwicky, A.M.: Speech acts distinctions in syntax. In: Shopen, T. (ed.) Language Typology and Syntactic Description, pp. 155–196. Cambridge University Press, Cambridge (1985)
9. Aikhenvald, AYu.: Imperatives and Commands. Oxford University Press, New York (2010)
10. Searle, J.R.: Expression & Meaning: Studies in the Theory of Speech Act. Cambridge University Press, Cambridge (1979)
11. Fortuin, E.L.J., Boogaart, R.J.U.: Imperative as conditional: from constructional to compositional semantics. Cogn. Linguist. **20**(4), 641–673 (2009)
12. Fortuin, E.L.J.: Polysemy or monosemy: Interpretation of the imperative and the dative-infinitive construction in Russian, Ph.D. thesis, Amsterdam, Institute for Logic, Language and Computation (2000). https://www.illc.uva.nl/Research/Publications/Dissertations/
13. Russkaya Grammatika, t.2: Sintaksis [The Russian Grammar, vol. 2: Syntax] (1980). http://rusgram.narod.ru/
14. Russian National Corpus. http://www.ruscorpora.ru/
15. Khrakovskij, V.S., Volodin, A.P.: Semantika i Tipologija Imperativa: Russkij Imperativ [Semantics and Typology of Imperative: Russian Imperative]. Nauka, Leningrad (1986)
16. Stojnova, N.: Pobuditel’nye upotreblenija form nastojashhego i budushhego vremeni [imperative use of forms of present and future tense]. In: Proceedings of the 3rd Conference «Issues of the language: The view of young scientists», pp. 227–243. Kanzler, Moscow (2014)
17. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. Linguist. **26**(3), 339–373 (2000)
18. Kral, P., Cerisara, Ch.: Dialogue act recognition approaches. Comput. Inform. **29**(2), 227–250 (2010)