Russian multiverb constructions

Natasha Akutina, *Linguistic Theory and Language Description*

I. Introduction

In Russian there are multiverb constructions that consist of two or three verbs, e.g. *idu kuryu* 'he/she is going (and) smoking', *sidit smejotsya* 'he/she is sitting (and) laughing', *idi l'ag polezhi* 'go lay down (and) lay' etc. In a sense, these constructions resemble serial verb construction structure since there is no 'overt marker of coordination, subordination, or syntactic dependency of any other sort' [Aikhenvald 2006]. Indeed, while pronouncing such a construction, speaker rather does not stop, thus it functions as a single predicate. Russian National Corpus data seems to support this idea, however only partly. The thing is that sometimes while writing speakers prefer to insert comma or hyphen between the units of a construction.

In this paper I will try to understand how crucial is the deviation between the two cases and how it influences our perception of such constructions as serial ones. My analysis is mostly based on the data found in the Internet, however, Russian National Corpus examples are also considered on the side. My research is restricted to constructions that can be found in the imperative form, such as *idi postoj podumaj* or *lyag polezhi* etc.

II. Russian 'serial' constructions

Russian multiverb constructions have been mostly studied by Weiss, who argues that they share all the main features to be considered serial. He works with his own sample of examples found in the Internet. For instance, he has an example *sjadem podumaem* 'we will sit down and think of it', but there are different variations to be found in the Internet, e.g. *sjadem podumaem* or *sjadem - podumaem or sjadem, podumaem*. Indeed, it depends on the context: e.g. if the speaker wants to highlight the process of thinking, probably he or she choses the third variant etc. The problem is that the absence of the comma can be interpreted as an ellipsis or even mistake.

Another problem with pseudo-serial constructions in Russian is that in serializing languages serial verb constructions serve the main part of cases. Very often, there is no other way for a speaker of a serializing language to describe a certain situation. There is no doubt that this is not true for Russian language. As a matter of formality I will look at the current situation with the distribution of examples in Russian National Corpus, however the number of examples is highly limited (that actually highlights the problem).

III. Data

Now I would like to describe my data. I took 17 constructions consisting of two and 9 constructions consisting of three verbs. They are:

- *l'ag pospi*; *l'ag polezhi*; *otkroju posmotrju*; *posizhu pochitaju*; *posizhu poslushaju*; *posidi podumaj*; *posidi otdohni*; *shodi uznaj*; *shozhu uznaju*; *shozhu sproshu*; *shodi posmotri*; *idi l'ag*; *idi shodi*; *idi posidi*; *pojdu otkroju*; *pojdu posizhu*; *pojdu shozhu*; *pojdi shodi*; *s'ad' posidi*;
- *idi l'ag pospi*; *idi shodi posmotri*; *idi l'ag polezhi*; *idi posidi podumaj*; *idi shodi uznaj; pojdu otkroju posmotrju*; *pojdu posizhu pochitaju*; *pojdu shozhu uznaju*; *pojdu shozhu sproshu; s'ad' posidi podumaj*; *s'ad' posidi otdohni*.

Firstly, let us discuss their structure. The starting point here is to analyse so called triplets, construction consisting of three verbs. It is easy to see that its first component can only be taken from a semantically restricted class of verbs, namely **verbs of motion** or **verbs of position**. It is important to notice, that it does not violate rules of serialization, as Aikhenvald notices: there are serializing languages that use only asymmetrical [1] type of serial constructions. These verbs are *idti/pojti*, *sidet'* (and technically *lezhat'*, e.g. *lyag polezhi posmotri televizor* 'lay down and lay and watch TV'). I will call first verbs, *idti* and *pojti*, M-verbs and in my data they are denoted as **M1**. Verbs like *sidet'* or *lezhat'* will be denoted as **P1** since they indicate position. Thus, my data for 'triplets' is marked as follows:

Table 1. Data example for triplets

construction	**type**	**V2**	**V3**	**IMP**
иди ляг поспи	M1	NO	NO	YES
иди ляг поспи	M1	NO	YES	YES
сядь посиди отдохни	P1	NO	NO	YES
иди ляг поспи	M1	YES	YES	YES
… etc.				

For example, the second row means that in the sentence there were no commas or hyphens before the second and the third verbs etc. Every construction is marked with M- or P-marker just in case that the type of the first row influences punctuation.

[1] **Asymmetrical** serial verb construction includes a verb from a grammatically or semantically restricted class; **Symmetrical** serial verb construction includes verbs only from unrestricted classes [Aikhenvald 2006].

It is important to say that there were few examples with hyphen that to some extent marks the strong connection between two (or three) verbs, but unfortunately we decided to remove such examples from the dataset since it adds noise to our difficult for the analysis data itself.

I had more than 400 marked sentences which can be found in my repository, however, the main research is made on the slightly cut data since even in the Internet the number of examples is restricted (however due to the aim of the research I need equal amount of sentences of each type).

Here I would like to describe briefly the procedure of the search, since sometimes it is really hard to find appropriate examples. Firstly, there are no repeated examples in the sample. Secondly, I was looking for examples written by native speakers who follow punctuation rules in the observable context. There are examples (1) that suits our sample pretty well and example (2) that does not:

(1) *Ну, иди сходи узнай про Трумена, а потом про Пол Пота поговорим.*

(2) *Мне гов**а**рят иди сходи узнай. Но я не люблю напрашиваться. Если бы они захотели бы меня взять на эту должность они бы позв**а**нили. Телефон есть.*

The second part of my data consists of the double verb constructions driven from the triplets discussed above. For example, if we have a triplet *idi shodi posmotri*, we have two double verb constructions driven – *idi shodi* and *shodi posmotri* etc. It is important to mark, that here we do not mark triplets context: context *idi shodi posmotri* cannot be counted for *idi shodi* as an example etc. The excerpt from the data is presented in Table 2:

Table 2. Data example for double verb constructions

construction	**type**	**V2**	**IMP**
иди ляг	M1	NO	YES
посижу почитаю	P2	NO	NO
сядь посиди	P1	NO	YES
сходи посмотри	M2	YES	YES
… etc.			

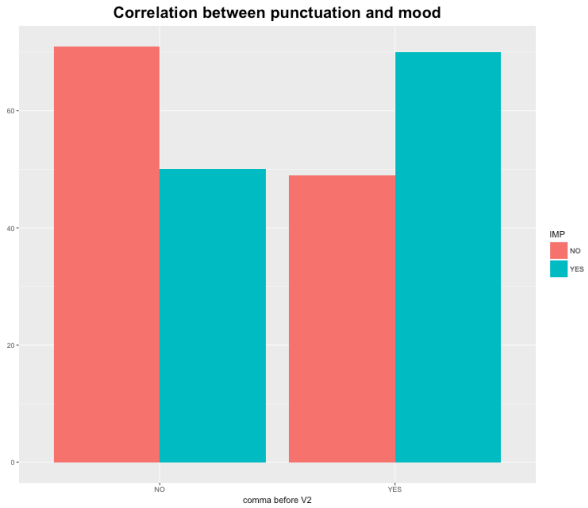
In Table 2 you can see new markers **M2** and **P2** that denotes *posidet'* as a position verb that **cannot** form a multiverb construction consisting of three verbs; or that *shodit'* is a motion

verb that also cannot form a triplet. This distinction between, e.g. *sidet'* (M1) and *posidet'* (M2), can be crucial[2].

After collecting the data from the Internet (Google search) I added another document that consists of the RNC examples, so it will be interesting to compare the results. Unfortunately, triplets are rather rare in Russian, that is why there is no evidence for chosen triplets in RNC (so I will compare only double verb results) and the amount of examples is partly limited for triplets found in Google (e.g. 20 sentences for *idi shodi uznaj* and 47 for *syad' posidi otdohni*, while for double constructions the amount per construction is fixed and is equal to 30).

IV. Correlations and crucial features (PART I)

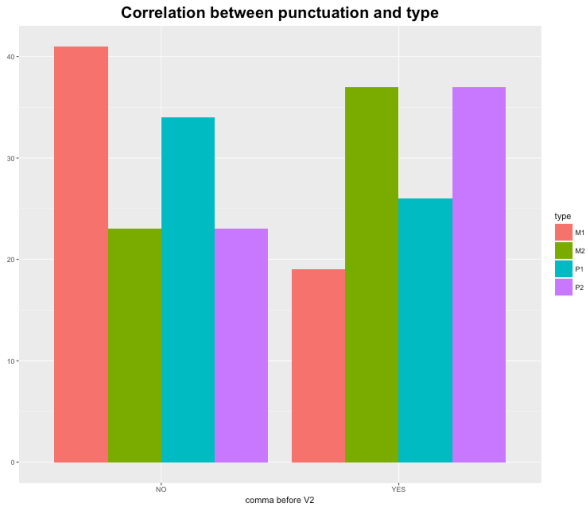
Let us visualize our data and look at some features separately. Since our data is mostly categorical and the amount of predictors is not big enough, we will see some correlations through visualization rather clearly. First of all, let us look at how the choice of comma correlates with the imperative/indicative form. For this purpose we take same constructions in different mood with four different types of the first verbs:



You can see that there is small correlation between the mood of the utterance and punctuation preferred by speakers. When the utterance is in the imperative mood the construction is likely to have comma between the verbs, but when the mood is indicative the absence of comma is preferable.

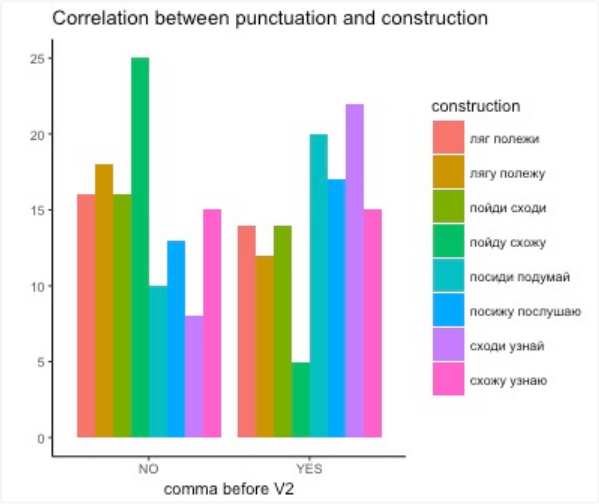
Secondly, I want to look at how the choice of position/motion verb correlates with the stops. For this purpose let us look at the same data:

[2] In principle, it seems like only *idti* or *poiti* can form strong triplets, however, if it is true, we will see it in our analysis anyway.



Here you can see that ‘strong’ motion verbs that can form triplets likely attach no comma, the same is with ‘strong’ position verbs. On the contrary, verbs that cannot form triplets are likely to be conjunct. Thus, so far we know that the mood of the utterance correlates with two factors: mood and semantic type of the verb.

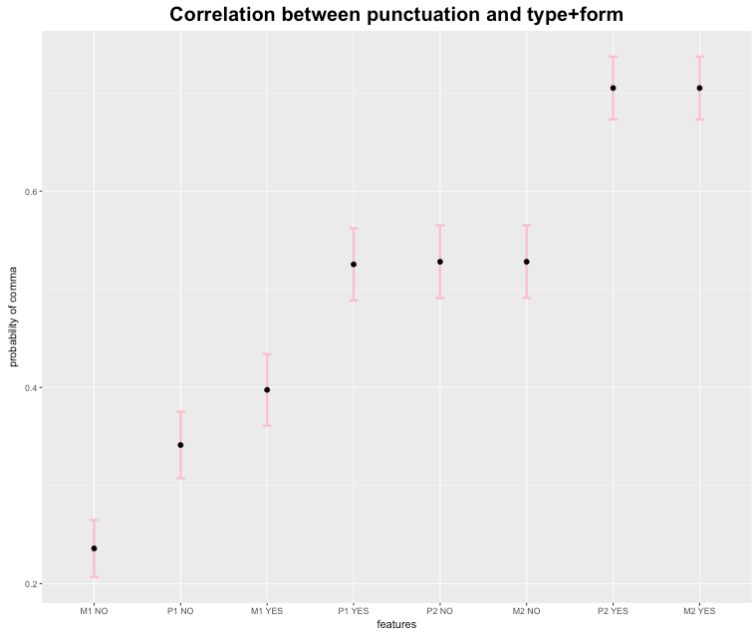
And here is a more detailed visualization based on separate constructions:



Now, let us fit Generalized Linear Model with our data and look at how our predictors work:

```
Call:
glm(formula = V2 ~ type + IMP, family = "binomial", data = y)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5630  -1.2209  -0.7335   1.1300   1.6997

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1756     0.3220  -3.651 0.000261 ***
typeM2        1.2882     0.3916   3.290 0.001003 **
typeP1        0.5178     0.3871   1.337 0.181069
typeP2        1.2882     0.3916   3.290 0.001003 **
IMPYES        0.7596     0.2729   2.784 0.005370 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.69  on 239  degrees of freedom
Residual deviance: 308.87  on 235  degrees of freedom
AIC: 318.87

Number of Fisher Scoring iterations: 4
```

We see that type and mood are to some extent significant factors. Now let us use *predict()* function to see the correlation more clearly. After giving the set of our features to it we can visualize the following model:



Here we can see that both predictors are relevant, hence the 'type' is more important. For example, we can see that M1 + indicative forms have 20% probability to attach comma, while M2 + imperative forms have nearly 80%.

V. Correlations and crucial features (PART II)

While searching for examples there was another interesting thing noticed. It seems that the sentences that are complete or full enough have a tendency to have comma. Consider the following examples:

(3) *Пойду схожу, посмотрю, кто пришел.*

(4) *Пойду сяду, посмотрю телевизор.*

(5) *Сядь, посиди во дворе.*

(6) *Сядь посиди.*

(5) *Пойду сяду посмотрю.*

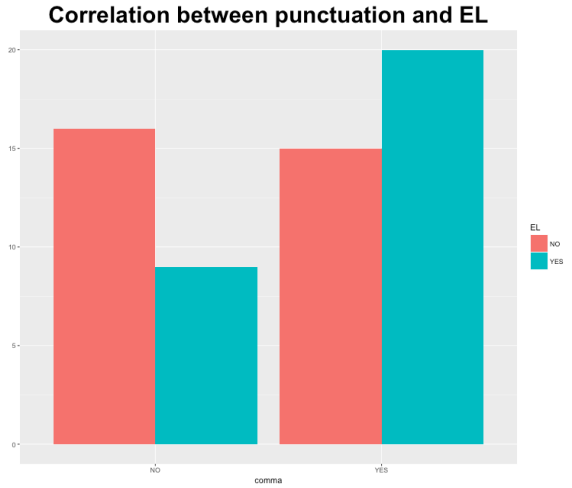
(7) *Пойду схожу посмотрю.*

If it is true, then we can probably postulate that the reason for such constructions being attested at all is ellipsis and it regulates the absence of a stop. Let us look at our new dataset for two constructions *shodi uznaj* and *shozhu uznaju*:

Table 3. Data example for double verb constructions

construction	**V2**	**ELL**
сходи узнай	YES	YES
схожу узнаю	YES	NO
…		

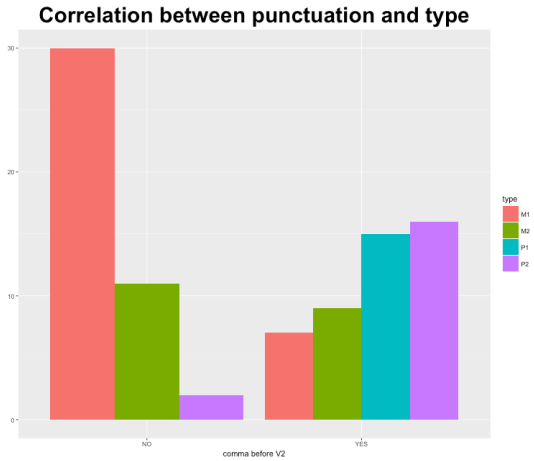
The main structure is the same, but the column ELL for ellipsis was added. Now we are going to visualize our new data:



Visualization shows that there is small correlation, however the p-value due to GLM is greater than 0.05.

VI. RNC

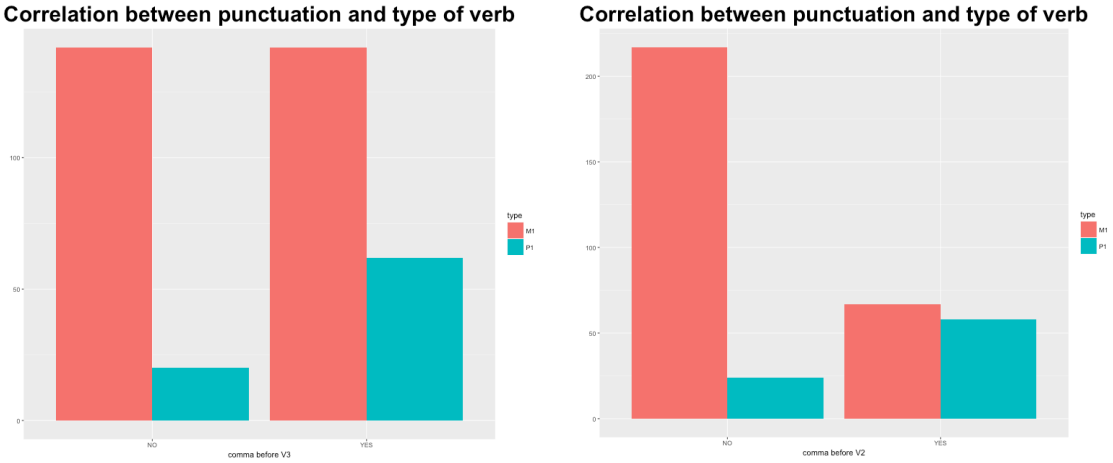
The main limitation of the dataset based on RNC examples is the unevenness of the sample (and the amount of examples at all). However, visualization gives us more or less the same result:



We can see that M1-verbs (like *idti*) rather do not attach comma, while P2-verbs (like *posidi*) are likely to have stops between its units.

VII. Triplets

After we have analyzed double verb constructions, let us see, are there any correlations between the semantic type of the first verb and stops within triplets:



There is strong correlation for M1-verbs and P2-verbs that show exactly the opposite results. It is important to mention here that we have some kind of hierarchy here: M1 – P1 – M2 – P2, so these results to some extent are predictable.

VIII. Conclusion

During my analysis I have checked which factors can determine the punctuation rules observed in so called serial constructions in Russian. I took Russian multiverb constructions consisting of three verbs and split them into double verb constructions in order to show that the first position verbs forming triplets (M1, P1) and other verbs forming only double verb constructions (M2, P2) behave differently.

The features we have discovered are:

1) **mood**: it seems that there is small correlation between the form of the utterance and the stops between the components of a construction, however due to our data it is insignificant; it does not matter whether the example is in the imperative or indicative form;
2) **ellipsis**: this feature was revealed during the current research and was checked on the restricted amount of data, and due to our data it is also insignificant feature; there are examples without the ellipsis having no commas and vice versa;
3) **first verb type**: the most significant feature that was revealed; it strongly correlates with the absence/presence of a stop in the utterance.

All in all, the most significant feature is the semantic type of the first verb forming a construction. So we can propose that semantic types of verbs involved in 'serialization' in Russian form scale from M1-verbs (verbs of motion that can form triplets, like *idti - idi posidi pochitaj*; they seems to add semantic redundancy [Weiss 2012]) to P2-verbs (verbs of position that cannot form triplets, e.g. *posidet' – posidi pochitaj*). If there is a double-verb construction formed by the M1 or P1 verbs, it more likely to have no comma between their components. It means that these verbs forms more close relations with verbs they attach and the construction is strongly perceived by speakers as a single predicate.

References

Aikhenvald 2006 – Aikhenvald A. Y. Serial verb constructions in typological perspective //Serial verb constructions: A cross-linguistic typology. – 2006. – С. 1-68.

Weiss 2012 – Weiss D. Verb serialization in northeast Europe: The case of Russian and its Finno-Ugric neighbors //Grammatical replication and borrowability in language contact. – 2012. – С. 611-646.

GitHub: https://github.com/akutina/rproject.data