Gerasimova Anastasia (anastasiagerasimova432@gmail.com)
Lyutikova Ekaterina (lyutikova2008@gmail.com)
Lomonosov Moscow State University
Pushkin State Russian Language Institute

INVESTIGATING MISMATCH BETWEEN PRODUCTION AND PERCEPTION OF VARIATION: AN EXPERIMENTAL STUDY•

1. Problem

Where does the gradience in judgement data come from?

(Phillips 2010, Sprouse, Schütze 2013, Sprouse 2015 among others)

gradience results from factors other than grammar that affect language processing

grammatical knowledge itself is gradient

↓ Two classes of syntactic architectures (Sprouse 2015): ↓

Binary-categorical theories: a sentence is either generated or not generated in grammar, gradience is dictated by non-syntactic cognitive systems.

Weighted-constraint theories: combination of different grammatical constraints leads to a range of grammaticality levels.

↓ Challenge: different effect sizes of acceptability differences. ↓

Binary-categorical theories: processing difficulties.

Weighted-constraint theories: different combinations of weighted constraints values.

⇒ How to estimate the weights of constraints?

1.1. Previous research

Linguists try to relate acceptability judgements to usage data.

Approach 1: grammatical knowledge is probabilistic and determinates both frequency of occurrence and acceptability ratings.

To formalize usage data gathering investigators use probabilistic language models that are fitted to the annotated corpus data in a supervised (Bresnan 2007) and unsupervised manner (Lau et al. 2017, Sprouse et al. 2018).

Bresnan (2007): linguistic manipulations with contextual predictors affect both probabilities and acceptability judgements in the same directions.

Language models can achieve good levels of accuracy in predicting the gradient data (Lau et al. 2017); but they demonstrate substantial loss in coverage of phenomena that is captured by categorical grammars Sprouse et al. (2018).

Approach 2: production data can be obtained using experimental design.

Verhoeven, Temme (2017): forced-choice task.
Bermel et al. (2017): fill-in-the-gap task with two possible options.

Results: the Likert scale and forced-choice task correlate.

• The study has been supported by RSF, project #18-18-00462 "Communicative-syntactic interface: typology and grammar" at the Pushkin State Russian Language Institute.

1.2. Explicit and implicit assumptions in previous studies

Assumption 1: corpus maintains grammatical constraints that are implied by speakers in rating tasks.

— How frequency of occurrence for a certain construction is related to a grammar of an individual?
The language community that provides production frequencies and an individual who provides ratings might possess non-similar grammars.

— Texts of which genre can be comparable to results of acceptability tasks in which speakers are asked to evaluate *naturalness* of the stimuli?

— How to interpret low frequency spectrum?

Assumption 2: forced-choice task can provide production data.

— Forced-choice should be considered a rating task (Sprouse, Schütze 2014).

Assumption 3: the analysis of alternations that are dependent on a set of contextual predictors[1].

— The distribution of predictors dictates the quantitative values for frequency of variants.
The ratings for a certain phenomenon are compared not directly to the distribution of this phenomenon but instead to the distribution of predictors that favor a certain value.

Assumption 4: the analysis of pairwise phenomena.

— The analysis of pairwise phenomena presupposes the binary distribution of language data: without violations and with violated functional or grammatical constraints.
The final comparison is performed between variables of different dimensions: binary choice in production and a scale of acceptability in perception.

Assumption 5: the usage of unsupervised language models.

— Unsupervised language models take into account all kinds of information that can be retrieved from corpus. This is not necessarily the information that humans obtain when they acquire and use language.

2. The current study

Hypothesis: If the grammatical knowledge is probabilistic, one would see consistent patterns in production and comprehension of a single speaker, *without* mediation of the collective language system of all the speakers or speakers from a certain community.

New data: To avoid the binary distribution of language data and predictors:
⇨ phenomena that supposedly exhibit free variation: although variants may tend towards certain contexts, neither of them seems to violate any constraint and be unacceptable in a certain context.

Research question: **How the grammatical options can be distributed in both production and perception domains of a single speaker?**

[1] Predictors constitute contexts that are plausible for one variant or another, which means that there are also contexts where one alternative is acceptable, while the other is not.

We approach the correspondence of production and perception data adopting an experimental design alternative to those used in previous research:

— Instead of using corpus we use production data obtained experimentally from respondents who are later asked to make judgements.
— Instead of pairwise phenomena we examine language variation. The phenomena that we examine include those that have more than two alternatives, so when gathering production data, we do not end up with the forced-choice task.
— Judgements are collected formally using the conditions and materials from the production experiment.
— We analyze behavior of each participant across the production and acceptability judgement experiments.

2.1. Three phenomena of variance in Russian

Case variation in nominalizations.
In nominalizations with lexically governed internal argument the external argument can be marked both GEN and INSTR. The case marking strategy depends on the amount of structure that is nominalized: thus, the adverbial / PP- modification increases the acceptability of INSTR (Pereltsvaig 2017, Pereltsvaig et al. 2018).

(1) *torgovlja evreev / evreyami skotom*
trading Jews.**GEN** / Jews.**INSTR** cattle.INSTR
'trading in cattle by Jews'

Gender mismatch.
Gender mismatch occurs in the context of masculine nouns that denote professional status of humans and refer to females. In NOM.SG these nouns can trigger both masculine and feminine agreement on attributive modifiers and past tense verbs (Pesetsky 2013, Lyutikova 2015).

(2) GRAMMATICAL AGREEMENT pattern: all agreeing constituents are masculine.

a. *nov-yj zubn-oj vrach prishel*
new-**M** dental-**M** doctor.**M** arrived-**M**

REFERENTIAL AGREEMENT: modifiers are masculine, the verb is feminine

b. nov-yj zubn-oj vrach prishl-a
new-**M** dental-**M** doctor.**M** arrived-**F**

REFERENTIAL ATTRIBUTIVE AGREEMENT: non-classifying adjectives and the verb are feminine.

c. nov-aya zubn-oj vrach prishl-a
new-**F** dental-**M** doctor.**M** arrived-**F**

ILL-FORMED pattern: non-classifying adjective is feminine but the verb is masculine.

d. * *nov-aya zubn-oj vrach prishel*
new-**F** dental-**M** doctor.**M** arrived-**M**

'new dental doctor arrived'

Case mismatch in paucal constructions.
In paucal constructions feminine nominalized adjectives and adjectives that modify feminine nouns can be marked both NOM and GEN. The case marking partially depends on the context of the paucal construction: NOM is preferred in the argumental (DP) position, GEN in quantificational (QP and PP) positions (Lyutikova 2015).

(3) a. *dve gornični-yje / gornični -yx*
two maid(FEM)-**NOM**.PL / maid(FEM)-**GEN**.PL
'two maids'

b. *tri dobr-yje / dobr -yx devushki*
three kind(FEM)-**NOM**.PL/ kind(FEM)-**GEN**.PL girls.F
'three kind girls'

(4) DP context. Agreement with predicate.

a. [*Dve gornični-yje / gornični -yx*] *ubirali nomer k priezdu gostya.*
two maid(FEM)-**NOM**.PL / maid(FEM)-**GEN**.PL did the room before guest's arrival.
'Two maids did the room before guest's arrival'

PP context. Comparative construction.

b. *Etot vypusk* ***na*** [*tri yark-ije / yark-ix kartinki*]
This issue is **PREP** three bright(FEM)-**NOM**.PL/ bright(FEM)-**GEN**.PL pictures.F
bogache, chem vcherashnii.
richer than yesterday's.
'This issue is three bright pictures richer than yesterday's.'

QP context. Impersonal predicate, no agreement.

c. *Na stole* ***ostalos'*** *[tri igral'n-yje / igral'n -yx karty]*
On the table **left.IMPRS.PST** three playing(FEM)-**NOM**.PL/ playing (FEM)-**GEN**.PL cards.F
'There were left three playing cards on the table'

The three phenomena differ with respect to the type of variation.

Case variation in nominalizations: variation can be manipulated by adding PP into the structure; when there is no PP, no predictors are distinguished.

Gender mismatch: no predictors are distinguished.

Case mismatch in paucal constructions: contextual predictors.

⇨ The choice of such phenomena allows us to replicate the choice of data from both types of previous studies: those that used data with annotated predictors, and those which used raw data.

3. Experiments

We conducted 6 experiments, 2 for each phenomenon:
— 3 production experiments: respondents were providing the case / agreement morphology;
— 3 AJ experiments: respondents were providing acceptability judgements using a 5-point Likert scale.

Participants:
— 106 self-reported native Russian speakers took part in production surveys (mean age: 21, SD 5.3; min 15, max 49; 82 females).
— 5 month later 57 speakers out of the 106 completed AJ surveys (mean age: 21, SD 4.7; min. 17, max. 37; 43 females).

The participants performed the task remotely, via the web-based software Google Forms.

3.1. Nominalization experiment

Materials. PRODUCTION.
One factor – TYPE OF NOMINALIZED VERBAL STEM:
— transitive stems with lexically governed internal argument (variation expected)
— unergatives (no variation expected, baseline conditions)
— unaccusatives (no variation expected, baseline conditions)

16 sets of target sentences (4 sentences per confiriton);
32 filler items (counterbalancing by Latin square design).

Task. Fill-in-the-blanks task: speakers were asked to generate arguments of nominalizations assigning cases that sounded most natural to them. As arguments they used words mentioned in previous context.

(5) V tot mesjac **armija** osvobodila **stolicu**, i osvobozhdenie _____ _____ sil'no podnjalo boevoj duh vseh soldat.

That month the army.NOM reconquered the capital.ACC, and reconquest ____ ____ lifted the martial spirit. (To fill in: of the capital by the army.)

3.2. Gender mismatch experiment

Materials. PRODUCTION.
One factor – COMBINATION OF ADNOMINALS
(determiners: possessive, demonstrative pronouns; high adjectives; low adjectives).

(6)

1. *det*	*high adj.*	*low adj.*	our hard-working executive **supervisor** organized
2. *det*	*high adj.*		our hard-working **supervisor** organized
3. *det*		*low adj.*	our executive **supervisor** organized
4. *det*			our **supervisor** organized
5.	*high adj.*	*low adj..*	hard-working executive **supervisor** organized
6.	*high adj.*		hard-working **supervisor** organized
7.		*low adj.*	executive **supervisor** organized
8.	(no adnominals)		**supervisor** organized

det = determiner (possessive/demonstrative pronoun)

16 sets of experimental sentences (8 combinations, 2 sentences per combination);
32 filler items contained nouns that unambiguously denote sex of the referent.

Task. The context explicitly indicated the gender of the human. The target clause contained the noun phrase and the verb in past tense with gaps instead of endings. The task was to fill-in the gaps.

(7) Vsju noch' Tane ne udalos' somknut' glaz: **nash_ otvetstvenn_ proektn_** menedzher **gotovil_** prezentaciju reklamnoj kampanii dlja radioholdinga.

All night long Tanya (name of a female) didn't have a chance to get a wink of sleep: **our responsible project** manager **was preparing** a presentation of promotional campaign for the radio corporation.

(8)

nash	otvetstvennyj	proektnyj	gotovila
our-M	responsible-M	project-M	was preparing-F

3.3. Paucal constructions experiment

Materials. PRODUCTION.
Three factors:
CONTEXT (QP, DP, PP)
ANIMACY (animate/inanimate)
PATTERN:
— paucal construction involves feminine nominalized adjectives;
— paucal construction involves modified feminine nouns.

24 sets of target sentences (12 conditions, 2 sentences per condition);
48 filler items.

Task. Provide case morphology for adjective and noun in a paucal construction that was given in brackets (the numeral was represented as a digit).

(9) _____ (2, prachechnaya) byli otremontirovany v etom mesyatse.
_____ (2, laundry(FEM)-NOM.SG) have been renovated this month.

3.4. Acceptability judgements experiments (one for each phenomenon)

Task. Evaluate the acceptability of sentences using a five-point Likert scale.

Materials. The number of conditions increased as we had to check acceptability for all possible variants that could be produced in the first experimental series.

Nominalization experiment: in each condition there was choice between GEN and INSTR
⇨ the number of stimuli multiplied by 2: 32 target sets.

Paucal construction experiment: in each condition there was choice between NOM and GEN.
⇨ the number of stimuli multiplied by 2: 32 target sets.

Gender mismatch experiment: for each combination there were four major patterns: grammatical agreement, attributive feminine agreement, referential agreement, and ill-formed agreement.
⇨ the number of stimuli multiplied by 4: 96 target sets.

4. Results

The theoretically predicted tendencies are supported by experimental data in both types of experiments. However, there is no ceiling effect for any variant in the target conditions in neither of the surveys.

(i) for the **nominalizations** derived from transitive stems with lexical government GEN is more frequent and more acceptable than INSTR;
(ii) the REFERENTIAL AGREEMENT pattern is the most frequent and the most acceptable choice for **gender mismatch** nouns;
(iii) in **paucal constructions** in argumental (DP) position NOM is more frequently used and is rated as more acceptable than GEN;
in **paucal constructions** in quantificational positions (PP and QP) NOM and GEN are both available and rated equally acceptable.

Do the results of the two types of experiments coincide?

Nominalizations: YES

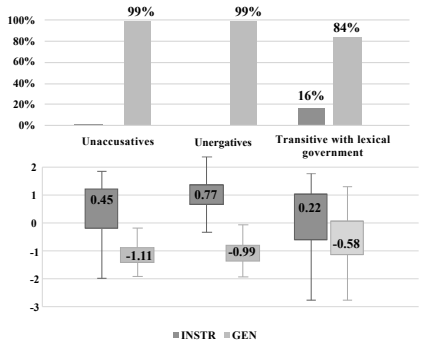

Figure 1. *Acceptability ratings and production experiment frequencies for the nominalizations*

Gender mismatch: NO

The two agreement patters, GRAMMATICAL AGREEMENT and REFERENTIAL ATTRIBUTIVE AGREEMENT, are produced and rated at different scales.

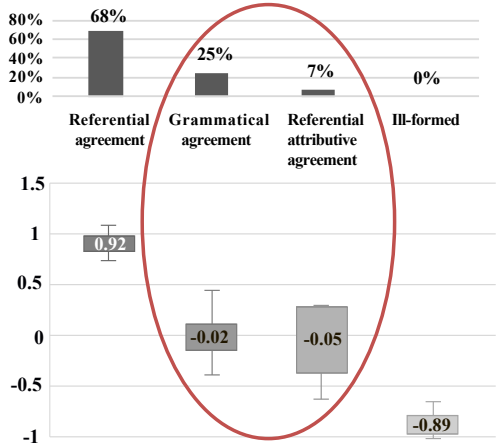

Figure 2. *Acceptability rating and production experiment frequencies for the gender mismatch patterns*

Paucal constructions: NO

In quantificational context (PP and QP) condition for adjectives there is a clear preference for GEN in usage and no preference in judgments.

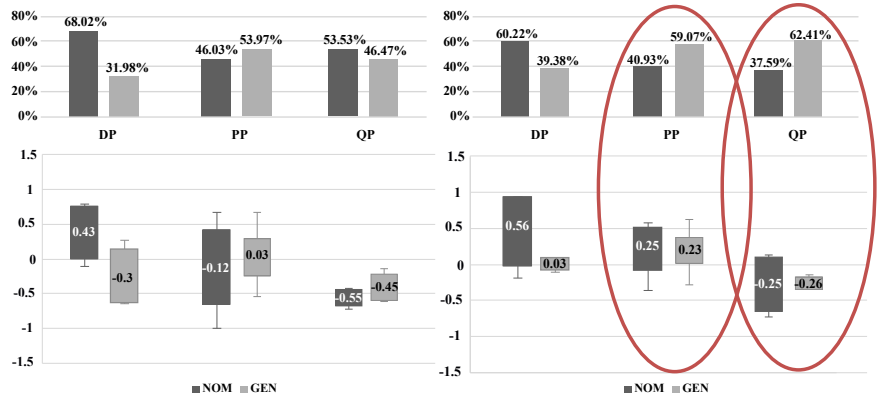

Figure 3. *Acceptability rating and production experiment frequencies for the paucal constructions with nominalized adjectives (on the left) and the paucal constructions with adjectives (on the right)*

4.1. Consistency of individual speakers

— The probabilistic nature of grammar presumes that frequencies of occurrence and acceptability scores are functions of the same grammatical constraints.
— Differences between the two modalities inevitably add noise and distortion to how the grammatical constraints are implemented

⇨ It is unreasonable to relate neither absolute nor proportional size of differences in ratings to the differences in frequencies.

⇨ We suggest analyzing whether acceptability direction is predicted by production or vice versa.

Does the respondent rate the variant that she used in the production experiment as more acceptable than the alternative?

For computation we registered:
(i) what the respondent has produced, one alternative or both, in a certain condition[2];
(ii) which of the two alternatives the respondent rated as more acceptable in the same condition.

Table 1. *Relative directional difference for the three experiments*[3]

Three strategies of choice and rating	Nominalizations	Gender mismatch[4]	Paucal constructions
What is produced is rated as the most acceptable	55%	57%	39%
In one experiment one out of the alternatives, in another – both	29%	30%	37%
Both variants in production	*25%*	*14%*	*23%*
Both variants in AJ	*4%*	*16%*	*14%*
Different alternatives in each experiment	16%	13%	24%

— Respondents stick to one variant only in half of conditions that allow for variation on the average.
— In nominalizations and paucal constructions experiments respondents were more likely to use both variants in production rather than in AJ experiment.
For gender mismatch experiments these rates do not differ.

In which experiment respondents are more consistent within one condition, production or AJ?

Within each experiment we analyzed whether a respondent was consistent across different lexicalizations of a single condition.

Table 2. *The consistency of respondents within one experiment*

	Nominalizations		Gender mismatch		Paucal constructions	
	Prod.	AJ	Prod.	AJ	Prod.	AJ
The same variant within one condition (is produced / rated as the most acceptable)	73%	94%	85%	82%	71%	80%
Different variants within one condition (are produced / rated as the most acceptable)	27%	6%	15%	18%	29%	20%

— In nominalizations and paucal constructions experiments there was more variability within production than in acceptability judgments.
— In gender mismatch experiments the variation was at equal rates.

⇨ Paucal constructions are the most unstable variation with respect to the two other phenomena.

[2] When making comparison only those conditions are taken into account that allowed for variance.
[3] In each table cell we present the percentage of cases, when the respondent demonstrated certain behavior towards an experimental condition. All conditions were taken from the production experiments.
[4] As there were more than one theoretically possible pairs of alternatives, we counted that 'both variants were used or rated acceptable' when one of these variants was used or rated acceptable in the other experiment.

5. Discussion

Three important findings:
1) data supports the idea of alignment between acceptability ratings and frequency of occurrence;
2) distribution of frequencies and acceptability scores do not coincide for all the alternatives;
3) different phenomena exhibit different values for consistency measures.

5.1. Inconsistency and the diachronical status of a phenomenon

The existence of several possibilities does not seem to be in accordance with the Economy principle. Either the alternation disappears, or the distribution of the variants becomes specified.
⇨ Inconsistency from one respondent can be expected.
Types of inconsistency differ, and variance can be further characterized from this point of view.

The degree of coherence of the two experiments reflects the effects of different stages of the variance evolvement.

— **Nominalizations:**
INSTR case marking is a rather new strategy (Pereltsvaig et al. 2018).
Due to the innovative nature, the strategy is still rated as unacceptable by those respondents who use it.

— **Gender mismatch:** The two alternatives are equally rated.
Muchnik (1971): questionnaire completed by 3780 speakers
GRAMMATICAL AGREEMENT in 38.6% of cases (cf. 25.21% in the current study)
⇨ Judgments reflect gradual decrease in production frequency of GRAMMATICAL AGREEMENT pattern in comparison to the leading pattern which is REFERENTIAL AGREEMENT.

— QP contexts in **paucal constructions**: The two alternatives are equally rated.
Specification of variants with respect to the structural position is predicted in accordance with the Economy Principle.

⇨ ***nominalizations:*** *development of the competing variant.*
⇨ ***gender mismatch and paucal constructions:*** *effects of different stages of the disappearance of variance.*

5.2. Inconsistency and the experimental methodology

— **Elicited production and AJs vary with respect to how they reveal variance in language.**
AJ in general show less variance; acceptability reaction is affected by other cognitive mechanisms that are involved into the process of decision making.

— **Production method as a more sensitive one?**
Neither production, nor AJ data provide a direct access to the grammar: they add distortion of different kind as different sets of cognitive systems are involved in the processes of production and perception.

— **Sensitivity of methods to different aspects of language phenomena:**
The elicited production is better in revealing deviations from the patterns prescribed in grammars;
the AJs are better at investigating to what extent a grammatical innovation is established in the language.

5.3. Inconsistency and the two grammar architectures

— **Are weighted-constraint theories wrong?**
The data shows that there is a clear correspondence between AJs and frequency of occurrence in half of conditions. However, inconsistency rates in general are far from being random.

The theories with both types of architecture, weighted-constraint and binary-categorical, do not necessarily take into account **the language redundancy**.
The variants that are at the low frequency spectrum might indeed be
(i) the residual effects of language evolvement or, the other way around
(ii) prerequisites for next changes.

Adding diachronical perspective could shed light on what is going inside the human mind synchronically.
The identified variance regularities should be considered not only when interpreting the data but also when modelling a language phenomenon.

5.4. Implications for methodology

— **The way the data sources conform can serve as an additional descriptive measure.**
In experimental methods the sample is analyzed as a whole, the individual differences are averaged out. The properties of individual behavior towards a certain phenomenon might provide a glimpse on its current state.

Summary:

- We investigated the correspondence between *offline production* and *offline perception* in the speech of a single speaker.
- The study focused on variance and examined three types of constructions that display a certain degree of variability.
- Data suggests that there is correspondence between frequency of occurrence and acceptability rates. However, this correspondence is more complicated than it was stated in previous studies.
- The way the two sources of data conform allows us to distinguish different types of variance, define unstable language domains, and, furthermore, can serve as an additional descriptive measure.

Appendix 1.

Table 3. *Summary of the studies that aimed at connecting production and AJ data.*

Study	**Phenomenon**	**Source of frequency data**	**Acceptability judgement task**	**Respondent samples for production and AJ data**	**Experiment timing**
Bresnan (2007)	English dative alternation	Switchboard corpus of spontaneous speech with annotated predictors	Forced-choice + confidence level	—	—
Lau, Clark, Lappin (2017)	500 sentences (BNC) 2000 sent. generated by round-trip machine translation	British National Corpus	Binary, 4-category and sliding scale	—	—
Sprouse et al. (in press)	Pairwise and multi-condition phenomena from LI and Adger (2003), 120 permutations *colorless green ideas* sentence	British National Corpus	Likert scale Forced-choice + Elo chess rating system	—	—
Verhoeven, Temme (2017)	SO and OS order in German clauses	Forced-choice	Likert scale	Same samples	Simultaneous
Klavan, Veisman (2017)	Estonian adessive case and adposition *peal* 'on' alternation	Morphologically Disambiguated Corpus of Estonian with annotated predictors	Forced-choice Likert scale	Different samples	—
Bermel et al. (2017)	Morphological variation in Czech case forms	Czech National Corpus Fill-in-the-gap task (restricted to forced-choice)	Likert scale	—	—
Current study	**Three phenomena that display a certain degree of variability**	**Production experiment with fill-in-the-gap task (not restricted to forced-choice)**	**Likert scale**	**Same samples**	**5 months in-between experiments**

References:

Adger D. (2003). Core syntax: A minimalist approach. Oxford: Oxford University Press, 2003.
Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in search of its evidential base, 96*, 77-96.
Bermel N., Knittl L., Russell J. (2017). Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguistics and Linguistic Theory.*
Divjak D. (2017). The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in Polish. *Cogn Sci, 41*, 354-382.
Klavan J., Veismann A. (2017). Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition peal 'on'. *ESUKA – JEFUL, 8(2),* 59-91.
Lau J. H., Clark A., Lappin S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cogn Sci, 41,* 1202-1241.
Lyutikova E. (2015). Features, agreement, and structure of the Russian noun phrase. *Russkii yazyk v nauchnom osveshchenii, 30,* 44–74.
Pereltsvaig A. (2017). Russian eventive nominalizations and universality of Determiner Phrase. *Rhema. Рема, 4,* 108-122.
Pereltsvaig A., Lyutikova E., Gerasimova A. (2018). Case marking in Russian eventive nominalizations: inherent vs. dependent case theory. *Russian Linguistics, 37(2),* 1-16.
Pesetsky D. (2013). Russian case morphology and the syntactic categories. Cambridge.
Phillips C. (2009). Should we impeach armchair linguists. *Japanese/Korean Linguistics*, 17, 49-64.
Schütze, C. T., & Sprouse, J. (2014). Judgment data. *Research methods in linguistics*, 27–50.
Sprouse, J. (2015). Three open questions in experimental syntax. *Linguistics Vanguard, 1(1),* 89–100
Sprouse J., Yankama B., Indurkhya S., Fong S., Berwick R.C. (in press). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review.*
Verhoeven E., Temme A. (2017). Word order acceptability and word order choice. *Linguistic Evidence 2016 Online Proceedings*. Tübingen.