Gerasimova Anastasia (anastasiagerasimova432@gmail.com)
Lyutikova Ekaterina (lyutikova2008@gmail.com)
Lomonosov Moscow State University
Moscow Pedagogical State University

COMBINING PRODUCTION AND JUDGMENTAL DATA: LANGUAGE VARIATION PERSPECTIVE[•]

1. Problem and previous approaches

Where does the gradience in judgement data come from? (Phillips (2010), Sprouse, Schütze (2014), Sprouse (2015) among others)

— gradience results from factors other than grammar that affect language processing;
— grammatical knowledge itself is gradient.

Sprouse (2015). Two classes of syntactic architectures correspond this opposition:

> **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 for both architectures: different sizes of acceptability differences.

> **Binary-categorical theories**: processing difficulties.
>
> **Weighted-constraint theories**: different combinations of weighted constraints values.

⇒ How the weights of constraints can be estimated?

1.1. Linking acceptability ratings and corpus data

Bresnan (2007): grammatical knowledge is inherently variable and probabilistic;
determinates both frequency of occurrence and acceptability ratings.

Materials: English dative alternation.

Method: Statistical model predicts the choice of dative construction based on several contextual predictors. Two experiments evaluated how ratings provided by speakers correspond to probabilities predicted by the model: speakers had to rate the naturalness of the alternatives in the given context by distributing 100 rating points over the two alternatives.

Results: acceptability judgements corresponded the corpus probability. Linguistic manipulations with contextual predictors affect both probabilities and acceptability judgements in the same directions.

Lau, Clark, Lappin (2017): prediction of acceptability judgements using unsupervised language models trained on raw text without any annotation or set predictors.

Materials: sentences retrieved from BNC and mappings of these sentences with errors introduced by round-trip machine translation.

Method: probabilistic language models augmented with acceptability measures that compensated for additional frequency factors.

Results: some models achieve good levels of accuracy in predicting the observed gradient data (acceptability measure scores correlate with mean human judgments).

[•] The study has been supported by Russian Scientific Foundation (RSF), project #16-18-02003 "Structure of meaning and its mapping into lexical and functional categories of Russian" at MPSU.

Sprouse et al. (in press): replication of experimental work by Lau et al. (2017) with enriched datasets.

Criticism: round-trip translations might not create grammatical oppositions that are made by syntacticians.

Materials: randomly selected samples of pairwise and multi-condition phenomena from Linguistic Inquiry and Adger's (2003) Core Syntax textbook + data set containing all 120 permutations of five words in Chomsky's *colorless green ideas* sentence.

Results: Although probabilistic models are able to capture the gradience in judgments of individual sentences, they demonstrate substantial loss in coverage of phenomena that is captured by categorical grammars and can be revealed in controlled syntactic experiments.

In all the studies listed above:

— Linguists try to relate acceptability judgements to usage data.
— The production data was retrieved from corpus.

Assumption: corpus maintains grammatical constraints that are implied by speakers in rating tasks because all the text entries were produced by speakers of the same language.

Limitations of the approach:

⇨ Does corpus correctly represent usage? What types of texts are represented in corpus?

> Bresnan (2007): the Switchboard corpus of spontaneous speech.
>
> Lau et al. (2017), Sprouse et al. (in press): this factor was not controlled.
>
> The problem: texts of certain genres might not be comparable to results of acceptability tasks in which speakers are asked to evaluate naturalness of the stimuli.

⇨ How to interpret low frequency spectrum?

> Divjak (2017): language model for *that*-clauses in Polish.
> Results: implicit probabilistic syntactic knowledge is based not on n-gram frequency, but rather on higher-order knowledge known as schemata or rules.

⇨ What type of data should be considered in a probabilistic language model?

- Predictors, distinguished by linguists.
  It is not clear whether one is able to distinguish all predictors that affect the final result.
- Unsupervised language models.
  These models take into account all kinds of information that can be retrieved from corpus, which is not necessarily the information that humans obtain when they acquire and use language.

1.2. Gathering production data differently

Verhoeven, Temme (2017): comparing preference ratings and frequencies of choice with a maximally controlled design with two experimental procedures.

Material: SO and OS order in German clauses.

Method: forced-choice task to evaluate results of production. Assumption: at some point in the production process, the speaker compares a set of alternative expressions and judges their relative appropriateness in a particular context.

> The problem: forced-choice should be considered a rating task (Sprouse, Schütze 2014).
> The results of combination of forced-choice task and Elo chess rating system correspond to results of Likert scale task (Sprouse et al. in press).

Results: the results of the two experiments correlate, although in particular contexts alternatives allow for more variation.

The probable cause for such finding is that speakers were accomplishing the same task, even though the speaker samples were different.

Klavan, Veisman (2017): what types of acceptability judgements should be used for comparison with probabilistic language models?

Material: the alternation between the Estonian adessive case construction and the postpositional construction with *peal* 'on'.

Method: the performance of a corpus-based language model against the results of two rating experiments: forced-choice and Likert scale.

Results: forced choice data compared to rating data provides a slightly better reflection of the corpus. However, it turned out that one of the alternatives retains high rating across the conditions of the predictors. Therefore, the two sources of data provide complementary information.

Bermel et al. (2017): the distribution of existing options in fill-in-the-gap and rating tasks, completed by respondents simultaneously.

Material: morphological variation in Czech case forms.

Methods: the results of fill-in-the-gap task served as production data.

Bermel et al. (2017) notice that in their case it was a forced-choice situation, as there were only two possible options in two environments.

Results: proportional frequency of forms in a balanced corpus is connected with their acceptability and with the frequency with which speakers select such forms.

The problem with the simultaneous performance of two tasks (within a single questionnaire): acceptability ratings could influence the production results and vice versa.

The second portion of studies:

— Formalization of production data gathering turned out to be comparison of different acceptability rating methods and their correspondence to probabilistic language models.

The type of phenomena that was analyzed:

Lau et al. (2017), Sprouse et al. (in press): pairwise phenomena (a minimal pair of an acceptable string and an unacceptable string).

The analysis of pairwise phenomena presupposes the binary distribution of language data: without violations and with violated functional or grammatical constraints.

Other mentioned studies: alternations that were dependent on a set of contextual predictors, distinguished and annotated by investigators in advance.

There were contexts in which one alternative was acceptable, while the other was not.

⇨ There is no such distribution for intralingual variation: although variants may tend towards certain contexts, neither of them seems to violate any constraint and be unacceptable in a certain context.

⇨According to weighted-constraint theories, in case of variation judgmental data also correlates with frequency in production.

Table 1. *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**

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

3. The current study

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.

3.1. 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.'

PP context. Distributive construction.
c. *Kazhdaya vypusknitsa mozhet priglasit'* ***po*** [*dve znakom-yje / znakom-yx*]
each graduate can invite **PREP** two friend (FEM)-NOM.PL / friend(FEM)-GEN.PL
'Each graduate can invite at most two acquaintances.'

QP context. Impersonal predicate, no agreement.
d. 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.

4. Experiments

In order to compare production and perception domains experimentally, we conducted the three production experiments, one for each phenomenon, in which respondents were asked to provide the case / agreement morphology, and the three AJ experiments, in which respondents were providing acceptability judgements using a 5-point Likert scale.

Participants: 106 speakers took part in production surveys; 5 month later 57 speakers out of the 106 completed AJ surveys (mean age: 21; min. age 17, max. age 37; 43 females, 14 males).

4.1. Production experiments

4.1.1. Nominalization experiment:

Materials.
4 types of nominalized verbal stems: unergatives, transitive stems with lexically governed internal argument, unaccusatives and transitives (as control).
16 experimental sentences (4 sentences per type)

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.)

4.1.2. Gender mismatch experiment:

Materials.
8 combinations of adnominals (determiners: possessive, demonstrative pronouns; high adjectives; low adjectives). All combinations used are listed in (6).
16 experimental sentences (8 combinations, 2 sentences per combination).

(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)

Task.
Respondents were asked to read a compound sentence:
The first clause provided context that explicitly indicated the gender of the human denoted by the subject in the second coordinate clause. The second clause contained the noun phrase and the verb in past tense with gaps instead of endings. Native speakers were asked to write the attributive modifiers and the verb with the endings in the textbox so that the sentence was complete.

(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

4.1.3. Paucal constructions experiment:

Materials.
Control for the context (QP, DP, PP), animacy and the pattern: whether paucal construction involves feminine nominalized adjectives or modified feminine nouns.
24 sentences (12 conditions, 2 sentences per condition).

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.

5. Acceptability judgements experiments

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

Materials.
The difference from the production experiment was that now 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
Paucal construction experiment: in each condition there was choice between NOM and GEN.
⇨ the number of stimuli was multiplied by 2.

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 was multiplied by 4.

6. Results

6.1. Nominalization experiments

Production:
— Both GEN and INSTR were available for transitive stems with lexical government (Transitive-LEX).
— For Transitive-LEX GEN is more frequent than INSTR.
— INSTR is rarely used with unergatives.

Acceptability judgements:
— INSTR is significantly more acceptable with stems with lexical government than with unergative stems (*Student's t-test, p = 0,03*).

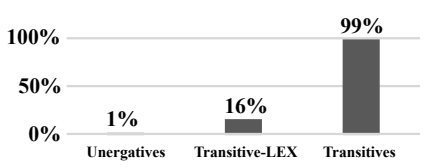

Fig. 1. Production experiment frequencies for the nominalizations with INSTR.

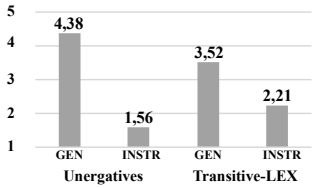

Fig.2. AJ rating for GEN and INSTR.

Speakers were inconsistent in using INSTR. Thus, the testee who constantly assign INSTR to external arguments of nominalization could estimate sentences with INSTR much less preferable than with GEN and vice versa.

6.2. Gender mismatch experiments

Neither the frequency nor the judgement of patterns for different combinations of adnominals differ significantly.

— Referential agreement is the most frequent and most acceptable pattern for all combinations.
— Referential agreement is significantly more acceptable than grammatical agreement and feminine attributive agreement (*Student's t-test, $p < 0.01$*).
— Although grammatical agreement and feminine attributive agreement had significantly different frequencies in the production experiment (25% vs. 7%), they had statistically equal acceptability scores (2.92 vs. 2.75).

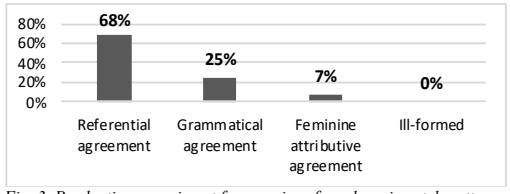

Fig. 3. Production experiment frequencies of gender mismatch patterns.

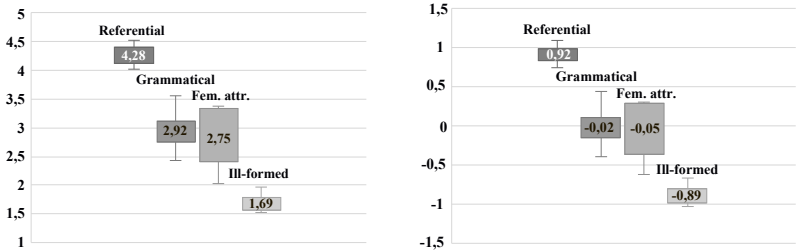

Fig. 4. Boxplot of AJ ratings for different patterns (raw data and z-score).

6.3. Paucal constructions experiments

Production:
— For nominalized adjectives in the argumental (DP) position NOM is preferred over GEN.
— For nominalized adjectives IN quantificational positions (PP and QP) both NOM and GEN are available.
— For adjectives in the argumental (DP) position NOM is preferred over GEN.
— For adjectives IN quantificational positions (PP and QP) GEN is preferred over NOM.

Acceptability judgements:
— For both nominalized adjectives and adjectives that modify feminine nouns in the argumental (DP) position NOM is rated as significantly more acceptable than GEN (*Student's t-test, $p < 0.01$*).
— For adjectives in the quantificational contexts (PP and QP) NOM and GEN have almost same acceptability ratings, although GEN is clearly preferred in production. The dispersion of the assigned acceptability scores for NOM is bigger than that of GEN in the same context.
— Sentences with QP contexts are in general rated lower than sentences with PP contexts.

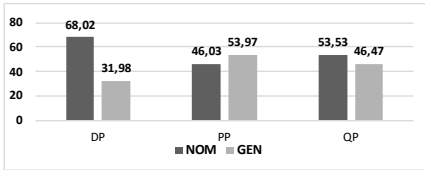

*Fig. 5. Production experiment frequencies for the paucal constructions with **nominalized adjectives**.*

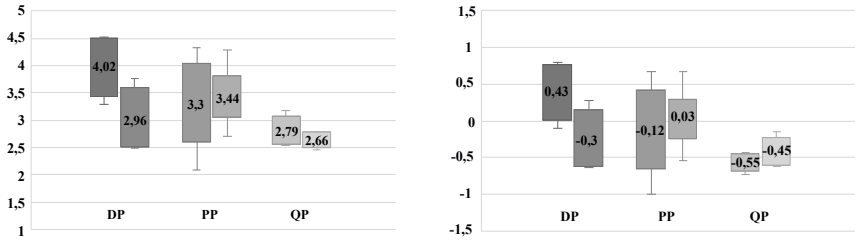

*Fig. 6. Boxplot of AJ ratings for the paucal constructions with **nominalized adjectives**.*

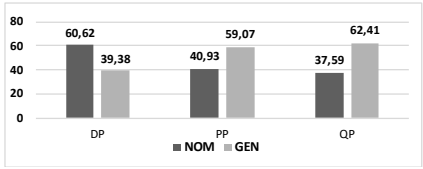

*Fig. 7. Production experiment frequencies for the paucal constructions with **adjectives**.*

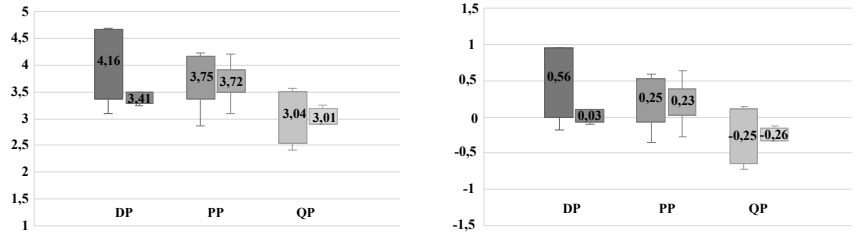

*Fig. 8. Boxplot of AJ ratings for the paucal constructions with **adjectives**.*

6.4. Relative directional difference

Throughout the three pairs of experiments the majority of respondents are inconsistent in production and perception.

Relative directional difference

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;
(ii) which of the two alternatives the respondent rated as more acceptable in the same condition.

Table 2. *Relative directional difference for the three experiments.*

Three strategies of choice and rating	Nominalizations	Gender mismatch	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%

⇨ 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.

Table 3. *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%

7. Discussion

Weighted-constraint theories suppose that acceptability ratings and frequency of occurrence are both functions of the same grammatical parameters. Our study shows, the results of production experiments do not correspond to the acceptability ratings in case of language variation, even when production and ratings are provided by the same respondents.

⇨ Where does the inconsistency come from?

The quality of production data

— Only one phenomenon presupposed binary distribution of answers: **nominalizations**, GEN or INSTR.
Gender mismatch experiment: respondents could choose from multiple variants, all of which were restricted to the phenomenon.

Paucal constructions experiment: respondents could choose alternative constructions: the interpretation of digits was not restricted, respondents could use collective numerals or quantificational nouns (5,33% of responses).

— Was **nominalizations** production experiment indeed a forced-choice task?
As forced-choice should be considered a rating task, we would not expect any differences in the results.

— The answers in our experiments could be more spontaneous. The enhancement of production data gathering could be a matter of future research.

The type of phenomena we examined

— 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.

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

— **Gender mismatch:** referential agreement is the main choice.
The two alternatives are equally tolerated. Grammatical agreement is still more frequent than feminine attributive agreement, but both variants have the same (rather low) level of acceptability.
⇨ Judgements reflect gradual decrease in production frequency of the grammatical agreement pattern in comparison to the leading pattern which is the referential agreement.

— QP contexts in **paucal constructions**: the respondents prefer one alternative, but rate both equally when perceiving them.
While there is a clear leader in production, judgements reveal this only partly – via the dispersion of possible answers.

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

Do our results suggest that weighted-constraint theories are wrong?

— The data shows that there are correlations.
— The problem is in the consistency of the respondents across different experimental methods. Inconsistency rates in general are far from being random.

Methodological source of results mismatch

— In different types of experiments respondents behave differently. Elicited production and acceptability judgments may vary with respect to how they reveal variance in language.

— AJ in general show less variance: speakers are put in a situation where they are forced to provide acceptability reaction that is affected by other cognitive mechanisms that are involved into the process of decision making. The task itself makes the choice more restricted.

— However, consistency rates are different in different series of experiments.

How our result can be extrapolated to other language domains that do not exhibit such variability?

— If we pick the production method as more sensitive, does it mean that we should base our models only on usage data? The choice of the experimental method depends on what we want our theory of language to model. Neither production, nor AJ data provide us a direct access to the grammar.

— The combination of production and AJ data allowed us to estimate the direction of changes in variability and see the full distribution of different variants. If we want our theory to capture constructions that are subjects to variance, we then should take into account both production and AJ data.

Summary:

- In the study we 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 combination of two sources of data provides the fuller description for cases of intralingual variation than a single method. The way the data sources conform allows us to distinguish different types of variance and, furthermore, define unstable language domains.

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.