UiT Norges arktiske universitet

Syntactic Singular vs. Semantic Plural in Russian, Ukrainian, and beyond

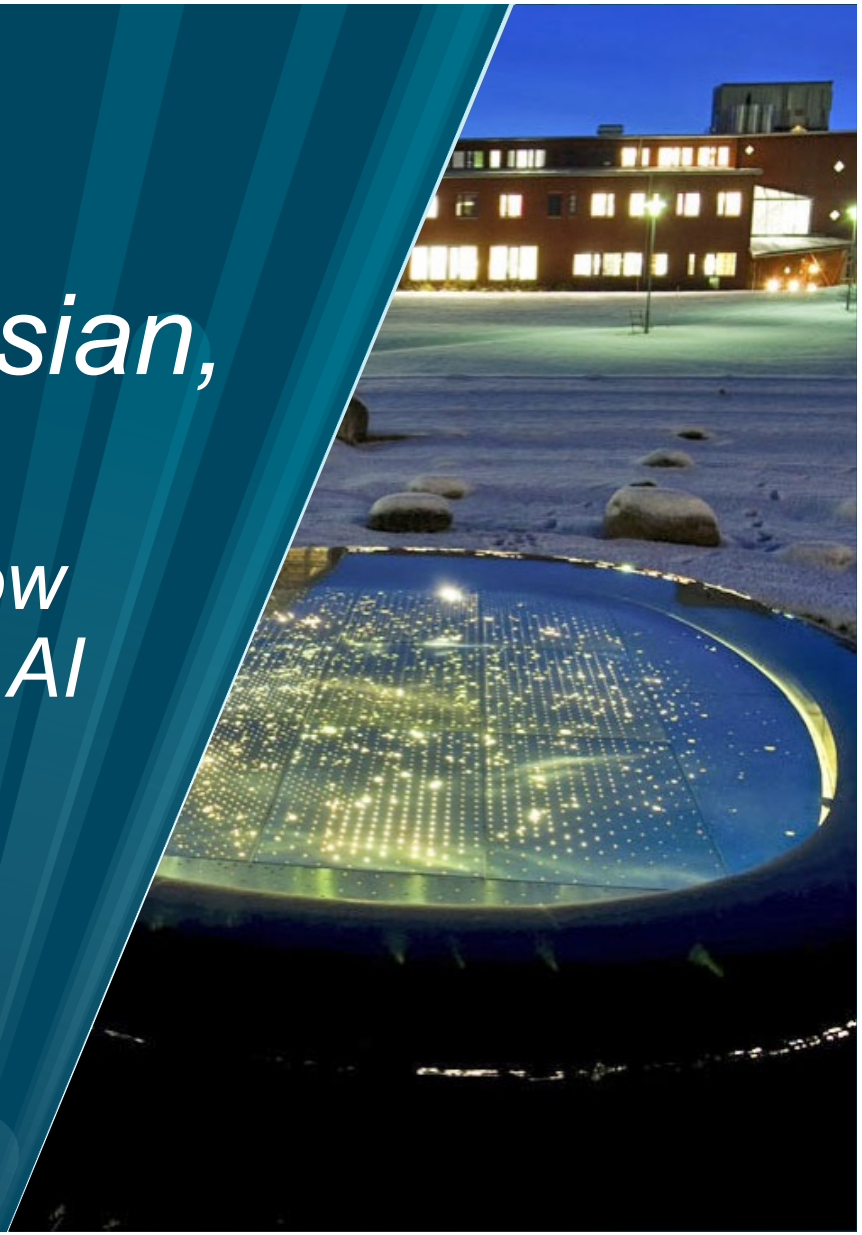
What is (not) happening and how that might change in the age of AI

Laura A. Janda

(with Tore Nesset and Yuliia Palii)

CLEAR

Cognitive Linguistics: Empirical Approaches to Russian



Overview

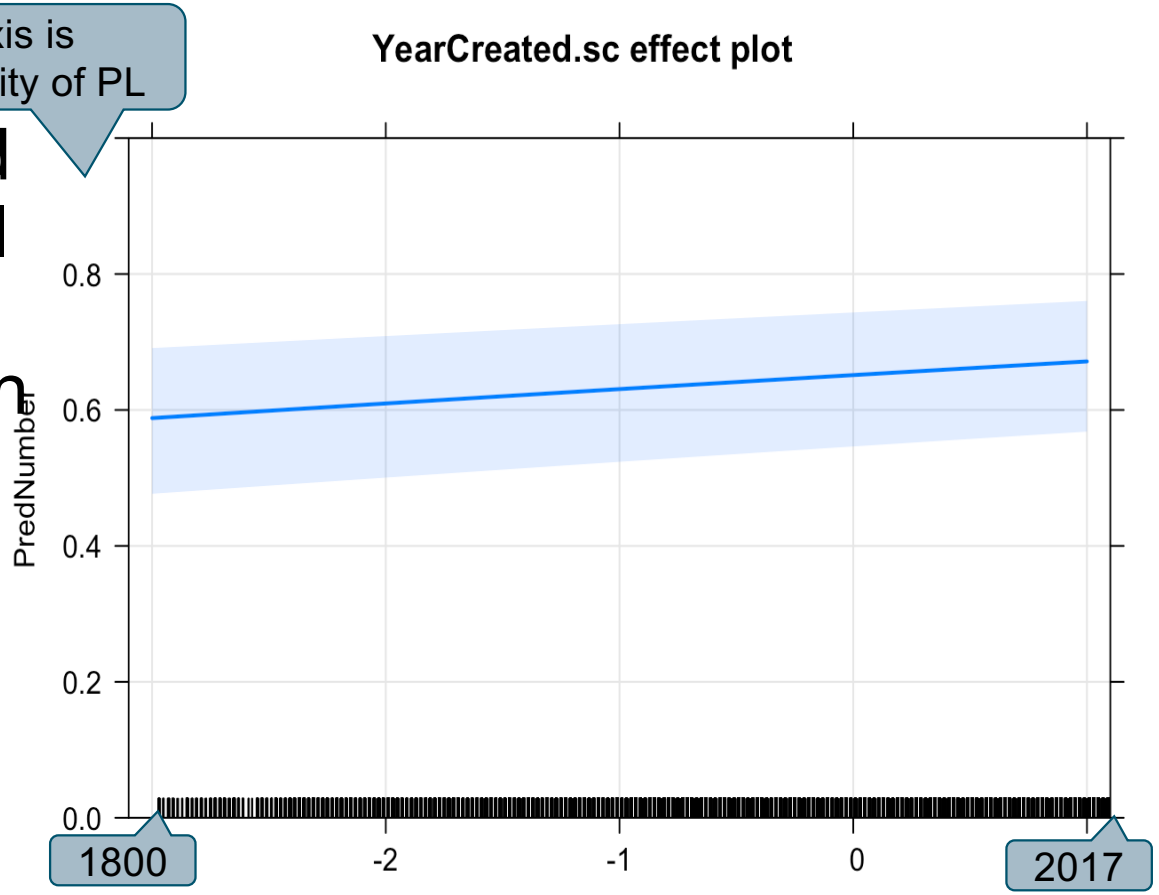
- How I got interested in this idea
  - Stability, variation, AI
- Singular vs. Plural variation
  - Alternative construals of the “same” reality
- Russian quantified subjects
- Ukrainian subjects quantified by *bahato*

UiT The Arctic University of Norway

How I got interested in this idea
Stability, variation, AI

How I got interested in this topic

Nesset & Janda (2023) found that Singular (44%) vs. Plural (56%) variation in verbs with quantified subjects in Russian (e.g. *ostalos'*[SG]/*ostalis'*[PL] *šest' čelovek* 'six people remained') has persisted nearly unchanged over the past 200 years



Surprising because Russian numeral morphosyntax has otherwise changed radically

The idea

Can LLMs impact the pace and direction of change?

- The environment
  - Stable co-existence of variant forms in language is common, though often overlooked
  - Preponderance of machine-generated language
- The issues
  - Do LLMs reproduce stable variation, or do they exaggerate biases?
  - If LLMs exaggerate biases, can they propel language change?

The one that got away…

- A grant proposal submitted to the Norwegian Research Council
- Missed funding by 1 point
- My university gave me a postdoc position as a “consolation prize”

Language variation

- When linguists look at variation, they tend to **focus more on change** than on stability
  - Exceptions: Longobardi 2001, Friðriksson 2008, Keenan 2009
- Linguists tend to **assume that variation must carry sociolinguistic indexical value** (Weinreich et al. 1968, Milroy 1992, Croft 2000, Blythe & Croft 2012)
  - Exceptions: Lass 1997, Ajión-Oliva and Serrano 2013

Spread of change

The spread of language change (“transition” cf. Weinreich et al. 1968) is facilitated by the presence of “numerous weak [social] ties”, in particular “first-order ties” realized as direct linguistic interactions (Milroy 1992, 2002) that can propagate change across a speech community.

In other words: language change spreads when **lots of people talk to one another**

Locus and quality of change

- Some have assumed that language change takes place **at L1 acquisition** (Andersen 1973 & 1989)
- Some have assumed that language change is **categorical**
- Others show evidence that language change takes place **across the lifespan** (Mechler & Buchstaller 2019)
- Others show evidence that language change is **gradual** (Bybee & Beckner 2015)

If the “others” are right, **adults** are liable to make **gradual changes** in language based on their exposure to **interactional use of language**.

My colleague in the Computer Science Department at UiT, Dilip Prasad

"Laura, most of what's on the internet is not going to be written by human beings."

Thompson et al. (2024) estimate that a "shocking" 57.1% of texts on the Internet are machine-generated, a growing trend, particularly for low-resource languages targeted by multi-way machine translations.

How LLMs work

- An LLM tokenizes language into units and computes the probability of the next unit in a string
- Bender et al. 2021 “On the Dangers of Stochastic Parrots”
- LLMs tend to perpetuate and often exaggerate biases present in training data (Manning 2022)

BUT: no one has looked at grammatical bias in LLMs

Human-AI interaction on a massive scale

- Millions (billions?) of people can converse with LLM-sourced chatbots on an unlimited and even intimate basis (cf. “numerous weak ties” and “first-order ties”)
- We know that human beings are sensitive to frequency distributions in language (Bybee 2015)
- Will LLMs serve the role of “sociolinguistic icons” (Eckert 2000) to which people accommodate their speech in a process of “koineization” (Kerswill 2002)?

UiT The Arctic University of Norway

Singular vs. Plural variation
Alternative construals of the "same" reality

Singular vs. Plural (SvP) variation

- Corbett (2000) describes number as at once deceptively simple, yet “the **most underestimated of grammatical categories**”, presenting complex facts that differ across languages.
- Grammatical number is the context for many types of variation, such as **syntactic Singular agreement** motivated by a Singular noun vs. **semantic Plural agreement** motivated by the fact that the noun refers to a multitude of individuals. This tug-of-war between syntax and semantics lies at the heart of controversies about grammatical agreement (Corbett 2006, Kibrik 2019).

Plurals and Masses in English grammar

Johnson (1987: 26) the same group of objects can be construed both as a mass and as a number of individuals

Lakoff (1987: 427–428) and Langacker (2008: 130–140):
masses portrayed as “non-plural mass” (*sand*) vs. “plural mass” (*cats*)

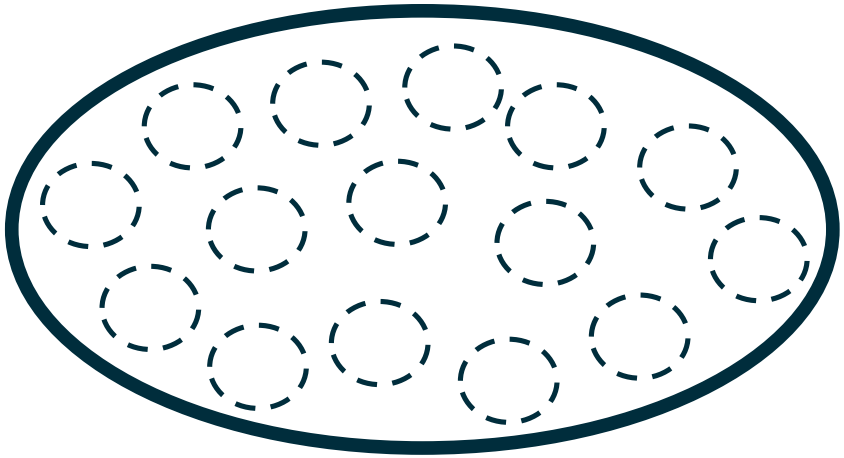
Mass and Plural behave the same way

(1) *all/most sand/cats/*cat*
(2) *They’re looking for *cat/sand/cats.*
(3) *a cat/*sand/*cats*
(4) *those/these/many/few/several/numerous cats vs. that/this/much/little sand*
(5) *three cats/*sand*

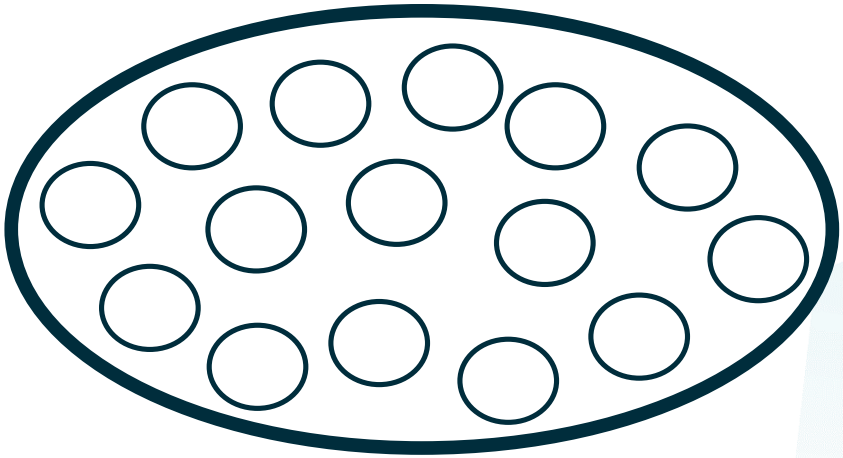
Alternative construals as either **mass** or **multiplex**

cf. Langacker (2008: 131), Johnson (1987: 26)

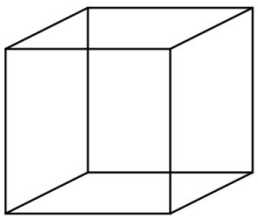
Non-plural Mass Noun

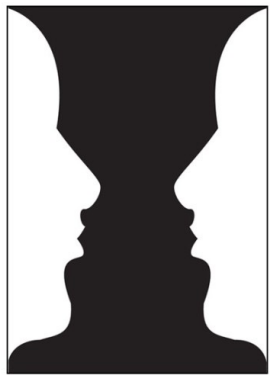


Plural Mass Noun



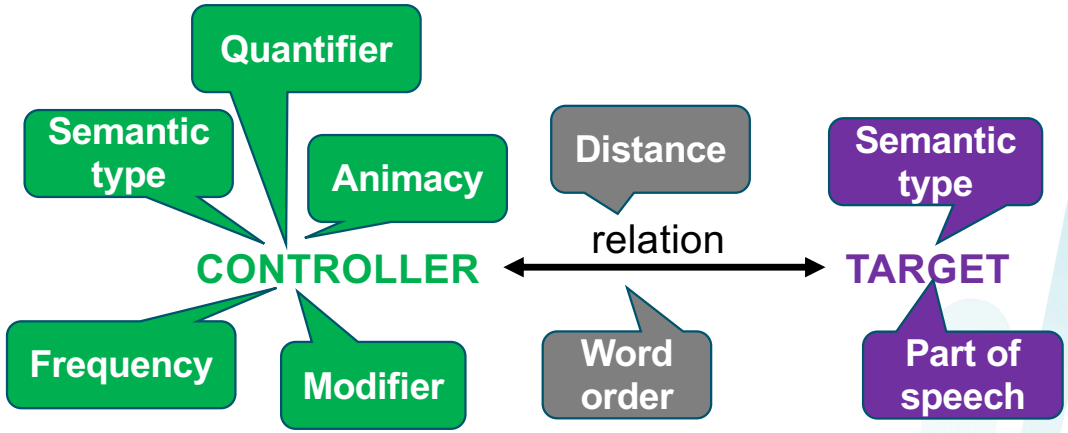
Cognitive multistability
(Stadler & Kruse 1995),
similar to ambiguous figures





Agreement, variation, learnability

1. In 1922, three years after the resumption of football, **Manchester United was** relegated to the Second Division.
2. At the end of the 1921–22 season, **Manchester United were** relegated to the Second Division, having won only eight games.



Agreement = “systematic covariance between a semantic or formal property of **one element** and a formal property of **another**” (Steele 1978)

UiT The Arctic University of Norway

Russian quantified subjects

DE GRUYTER MOUTON Cognitive Linguistics 2023; aop

Tore Nesset* and Laura A. Janda

A network of allostructions: quantified subject constructions in Russian

https://doi.org/10.1515/cog-2021-0117
Received November 6, 2021; accepted January 28, 2023; published online February 22, 2023

Abstract: This article contributes to Construction Grammar, historical linguistics, and Russian linguistics through an in-depth corpus study of predicate agreement in constructions with quantified subjects. Statistical analysis of approximately 39,000 corpus examples indicates that these constructions constitute a network of constructions ("allostructions") with various preferences for singular or plural agreement. Factors pull in different directions, and we observe a relatively stable situation in the face of variation. We present an analysis of a multidimensional network of allostructions in Russian, thus contributing to our understanding of allostructional

Case Study 2: Russian quantified NP + verb ‘Six persons remained.SG/remained.PL’

V okončatel’n-om sostav-e bratstv-a
[in final-M.LOC.SG set-LOC.SG brotherhood-GEN.SG

osta-l-o-s’ SG
remain-PST-N-REFL

šest’ **čelovek.** (Zaxarov 1988–2000)
six.NOM person.GEN.PL]
‘In the end six persons remained in the brotherhood.’

Quantified NP

Posle ix uxod-a na l’din-e
[after their departure-GEN.SG on ice.block-LOC.SG

osta-l-i-s’ PL
remain-PST-PL-REFL

šest’ **čelovek** […]. (*Znanie-Sila* 1997)
six.NOM person.GEN.PL]
‘After their departure six persons remained on the block of ice.’

Data:
- 38,988 examples from Russian National Corpus
- 36,182 examples after removal of nominative premodifiers

Baseline:
- 56% PL

Factors:
- Nominative premodifier
- Year Created
- Quantifier Frequency
- Quantifier Type
- Animacy
- Word Order
- Quantifier Lemma
- Verb Lemma

Nominative premodifiers: "**These** two professors gave a talk"

(2a) **Èt-i** **pjat'** **predloženij** da-l-i-s' mne tjažel-ee,
[this-NOM.PL five.NOM sentence.GEN.PL give-PST-PL-REFL I.DAT difficult-CMP

čem ves' ostal'n-oj tekst. (*Russkij reporter* 2013)
than whole.M.NOM.SG remaining-M.NOM.SG text.NOM.SG]
'These five sentences were harder to process than all the rest of the text.'

(2b) [.. **cel-ye** **tri** **knig-i** okazyva-jut-sja
[whole-NOM.PL three.NOM book-GEN.SG turn.out.to.be-PRS.3PL-REFL

pod zapret-om. (Rassadin 2004–2008)
under prohibition-INS.SG]
'As many as three books turn out to be forbidden.'

NomPs >> Pl in the literature:
- Crockett 1976
- Rozental' and Telenkova 1976
- Švedova 1980
- Kuz'minova 2004
- Timberlake 2004
- Pereltsvaig 2006

From 38,988 examples,
2,806 have NomPs:
2,801 have PL verb
5 have SG verb!

Since this is a (nearly) categorical rule, this factor was excluded from statistical model

- Confirmation of hypothesis that NomPs >> PL
- New: Cue reliability = 99.8%, but cue availability = 7.2% and cue validity < 7.2%
- Learnability: How are such rules learned?

Things to wrap your head around concerning numerals 2-900...

- Q: Which numerals are more likely to support a plural interpretation, and which ones are more likely to support singular?
- A: Smaller numerals allow for more individuation, thus supporting more plural, whereas large numerals are more indicative of masses that are perceived as wholes and therefore singular
- Q: Which numerals are of high frequency and which are of low frequency?
- A: Small numerals (2, 3, 4, etc.) are of higher frequency, bigger numerals have progressively lower frequency

- **Larger** numerals are of **lower** frequency and tend to have **singular** agreement.
- **Smaller** numerals are of **higher** frequency and tend to have **plural** agreement.

Mixed effects logistic regression model for predicting Predicate Number

PredNumber ~ 1 + YearCreated.sc + LogQuantFreq.sc + QuanType + Animacy * WordOrder + (1|Quantifier) + (1|VerbLemma)

“Predicate Number is predicted in relation to an overall intercept (1), with main effects of Year Created, Quantifier Frequency, and Quantifier Type, an interaction between Animacy and Word Order, and random effects of Quantifier and Verb Lemma”

Correct predictions for 84.2% of the data, well above the baseline of 55.96%; C score is 0.922

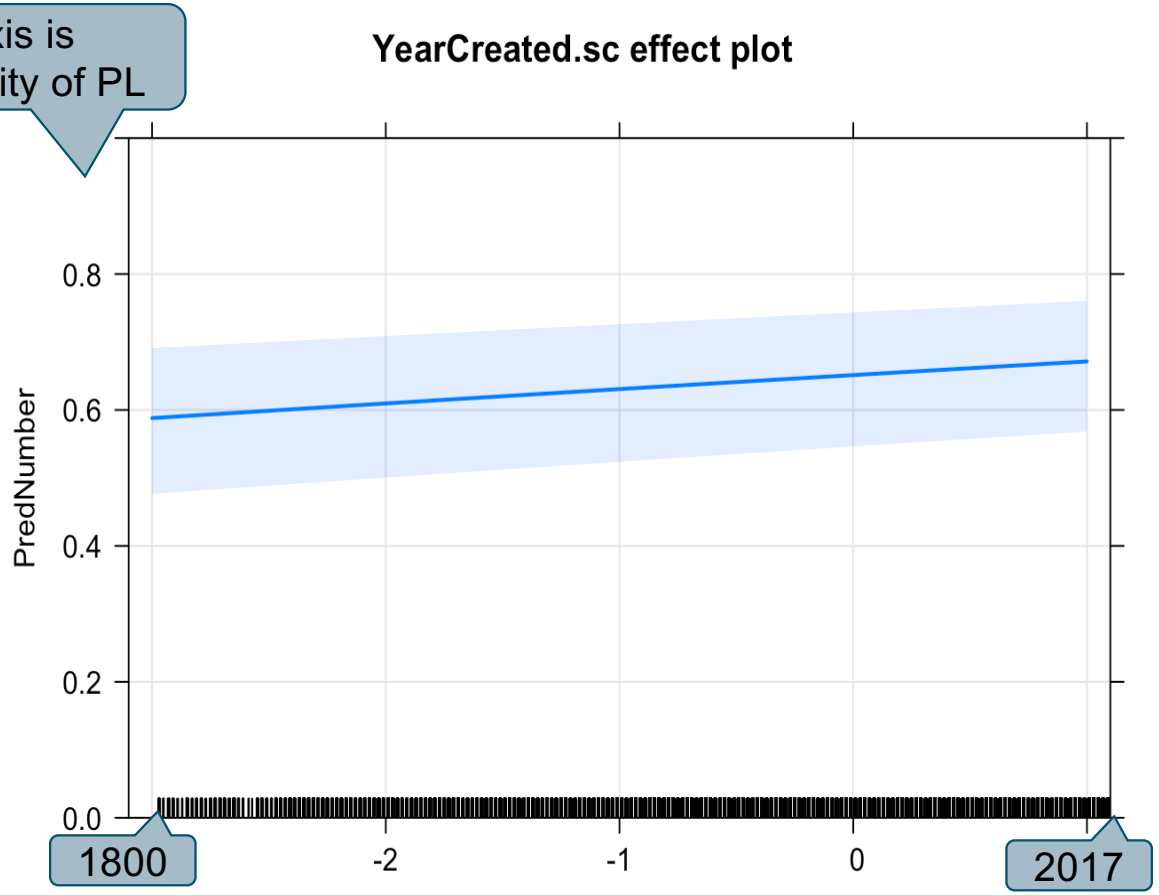
Results of logistic regression model for fixed effects

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.20724	0.30691	3.933	8.37e-05	***
YearCreated.sc	0.08970	0.01610	5.571	2.53e-08	***
LogQuantFreq.sc	0.39513	0.11103	3.559	0.000373	***
QuanTypecollective	-0.88051	0.46554	-1.891	0.058576	.
QuanTypenumeral	-2.03981	0.34382	-5.933	2.98e-09	***
QuanTypeindefinite	-4.60623	0.41864	-11.003	< 2e-16	***
Animacyanim	0.98308	0.04317	22.771	< 2e-16	***
WordOrderSV	1.28468	0.04896	26.239	< 2e-16	***
Animacyanim:WordOrderSV	0.59289	0.07251	8.177	2.91e-16	***

- A number of factors influence the choice between SG and PI agreement.
- One interaction is detected: Animacy and word order.

Language change? Year Created

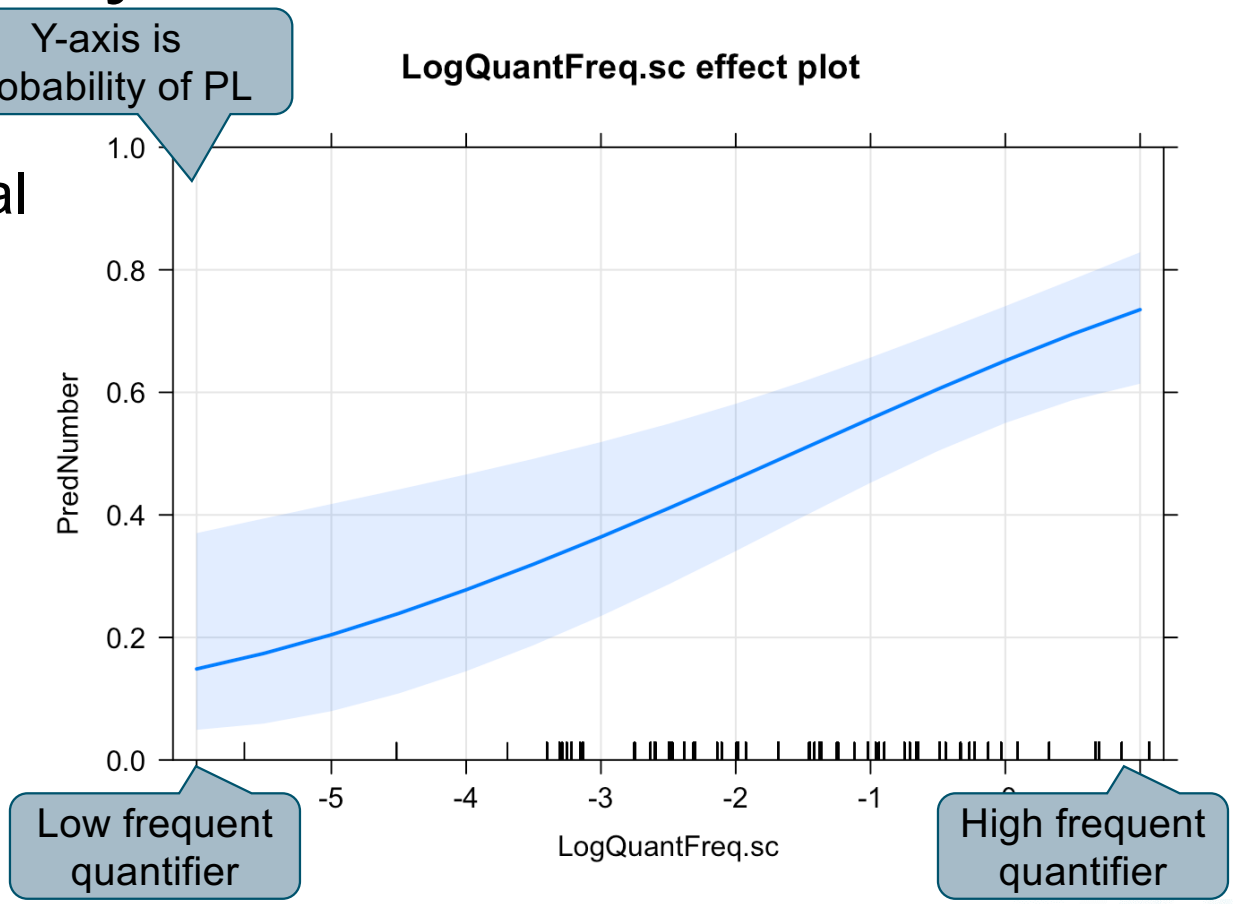
- Year of origin scaled (z-scored)
  - 1800 = -2.97 (minimum)
  - 1949 = 0 (mean)
  - 2017 = 1.37 (maximum)
- Some claim plural agreement has increased:
  - Gorbačevič 1971
  - Rozental' 1974
- Others find no evidence of change:
  - Corbett 1981



- Confirmation of hypothesis that situation is stable over time: "Static variation"
- Surprising: Russian numeral morphosyntax has otherwise changed radically.

Quantifier Frequency

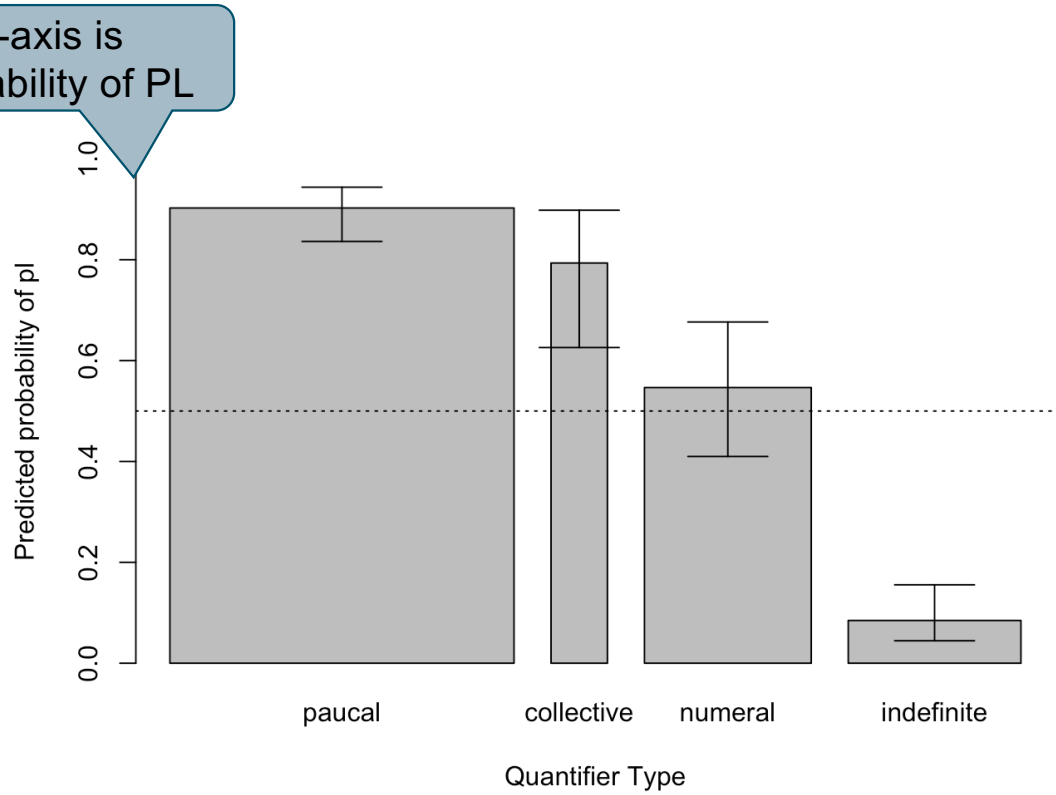
- Quantifiers have different frequencies in Russian National Corpus:
  - *vosem'sot* '800': 1,802 attestations
  - *dva* '2': 316,172 attestations
- Corpus frequency of each quantifier logarithmically transformed and scaled (z-scored):
  - 97 = -5.65 (minimum)
  - 159,318 = 0 (mean)
  - 352,781 = 1.07 (maximum)



PL agreement is more likely to occur if the quantifier is of high frequency

Quantifier Type

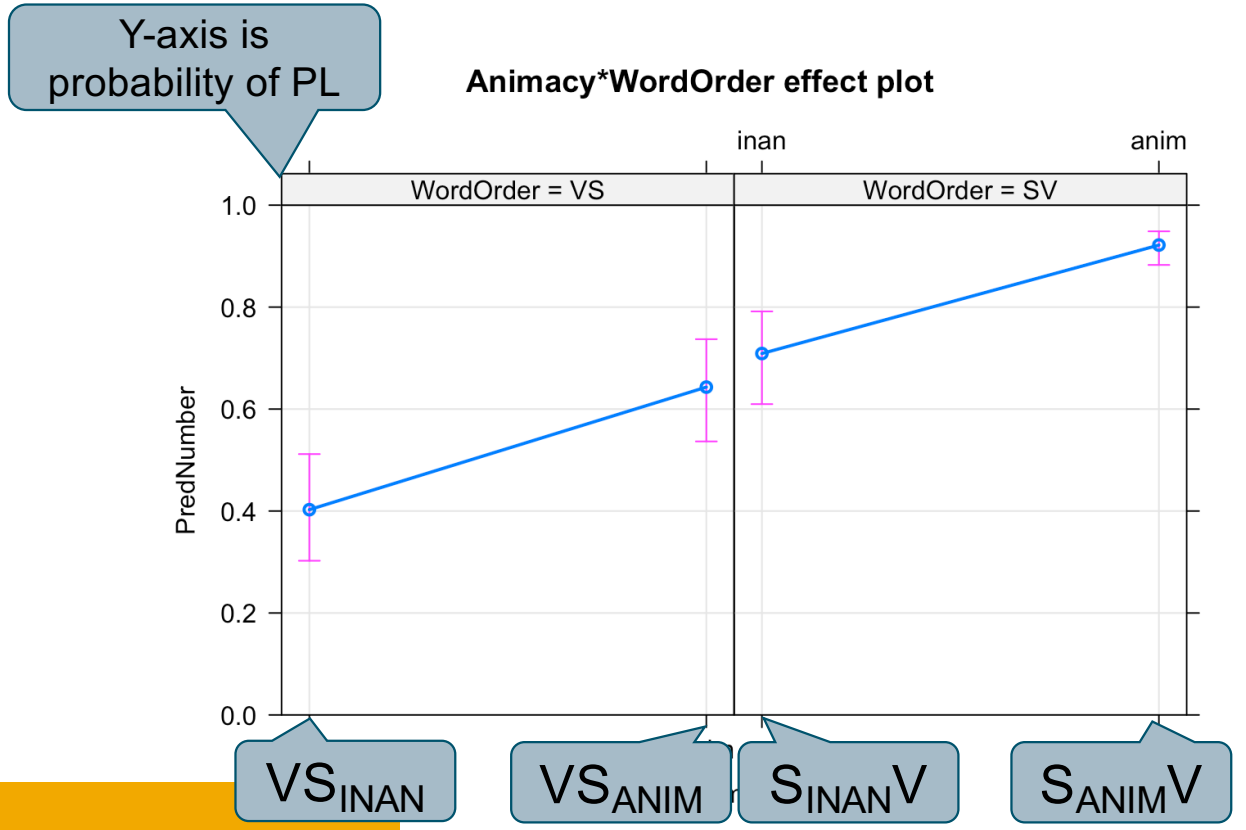
- Paucal: ‘2’, ‘3’, ‘4’, ‘both’, ‘one and a half’
  - stem etymologically from adjectives; NP has unique syntax
  - considered distinct by Corbett 1993; Pereltsvaig 2010; Igartua and Madariaga 2018
- Collective: ‘twosome’, ‘threesome’, ... ‘tensome’
  - used with human and pluralia tantum referents; NP in Genitive Plural
- Numeral: ‘5’ and above
  - stem etymologically from nouns; NP in Genitive Plural
- Indefinite: ‘many’, ‘some’, ‘how many’, ...
  - NP in Genitive Plural



Likelihood of PL agreement:
Paucal >> collective >> numeral >> indefinite

Interaction of Animacy and Word Order

- Regression model finds interaction between animacy and word order
- Four possible combinations
  - $S_{ANIM}V$
  - $S_{INAN}V$
  - VS_{ANIM}
  - VS_{INAN}



Likelihood of PL agreement:
$S_{ANIM}V$ >> $S_{INAN}V$ / VS_{ANIM} >> VS_{INAN}

Random effects 1: Quantifier lemmas

- Four types of quantifiers are important:
  - Paucal, collective, numeral, indefinite
- Do individual lexemes within each group behave differently?
  - If so, are there any patterns?

- **Case 1: Paucals ranked**

Definite: 'both' ≈ 'the two'

- *obe* 'both' (fem)
- *oba* 'both' (masc/neut)
- *dve* 'two' (fem)
- *četyre* 'four'
- *dva* 'two' (masc/neut)
- *tri* 'three'
- *poltory* 'one and a half' (fem)
- *poltora* 'one and a half' (masc/neut)

Definiteness effect: Definite >> other

FEM >> MASC/NEUT

- **Case 2: Indefinites ranked**

- *neskol'ko* 'some'
- *nemnogo* 'not many'
- *nemalo* 'not few'
- *malo* 'few'
- *stol'ko* 'so many'
- *mnogo* 'many'
- *skol'ko* 'how many'

Behaves like a numeral

- Definite promotes PL agreement (paucals)
- Feminine promotes PL agreement (paucals)
- Lexical items may show idiosyncratic behavior (*neskol'ko*)

Random effects 2: Verb lemmas

- Do individual verbs have different preferences?
  - Investigated lemmas attested over 100 times (57% of total), remainder listed as NotLemmatized
- Case 1: Semantic types of predicates
  - Robblee 1993:
    Agentive (e.g. 'work') >> Intransitive (e.g. 'be located') >> "Inversion" (e.g. 'be')
  - Our data provide some support: Agentive predicates generally promote PL agreement.
  - Situation may be more complex than Robblee anticipated.
- Case 2: Constructions
  - *Emu ispolnitsja**SG** pjat' let.* 'He turns**SG** five years old.'
  - *Prošlo**SG** dve nedeli*. 'Two weeks went**SG** by.'
  - *Na remont ušlo**SG** dva milliona dollarov*. 'Two million dollars were spent**SG** on the renovation.'
  - Measurement of time or other resource: always SG agreement (regardless of semantics of verb)

- Semantic types of predicates are relevant:
  - Agentive predicates promotes PL
- Constructions are relevant
  - Measurement of time/resource promotes SG

Russian quantified subjects: summary

- Premodified quantified subjects prefer PL (nearly categorical rule)
- Relatively stable variation 44% SG vs. 56% PL
- Low frequent (large) quantifiers prefer SG, High frequent (small) quantifiers prefer SG
- Paucals and collectives prefer PL, indefinites prefer SG, others about 50-50
- Animate SV prefers PL, VS inanimate prefers SG
- Quantifiers and verbs have individual preferences
- Measure constructions prefer SG

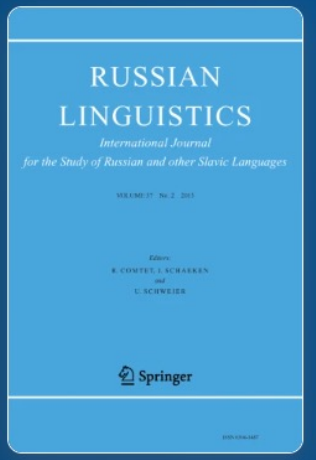
UiT The Arctic University of Norway

Ukrainian subjects quantified by *bahato*

Understanding 'many' through the lens of Ukrainian *багато*

Open access | Published: 17 October 2024

Volume 48, article number 18, (2024) Cite this article



Russian Linguistics

Laura A. Janda & Yuliia Palii

https://link.springer.com/article/10.1007/s11185-024-09301-7

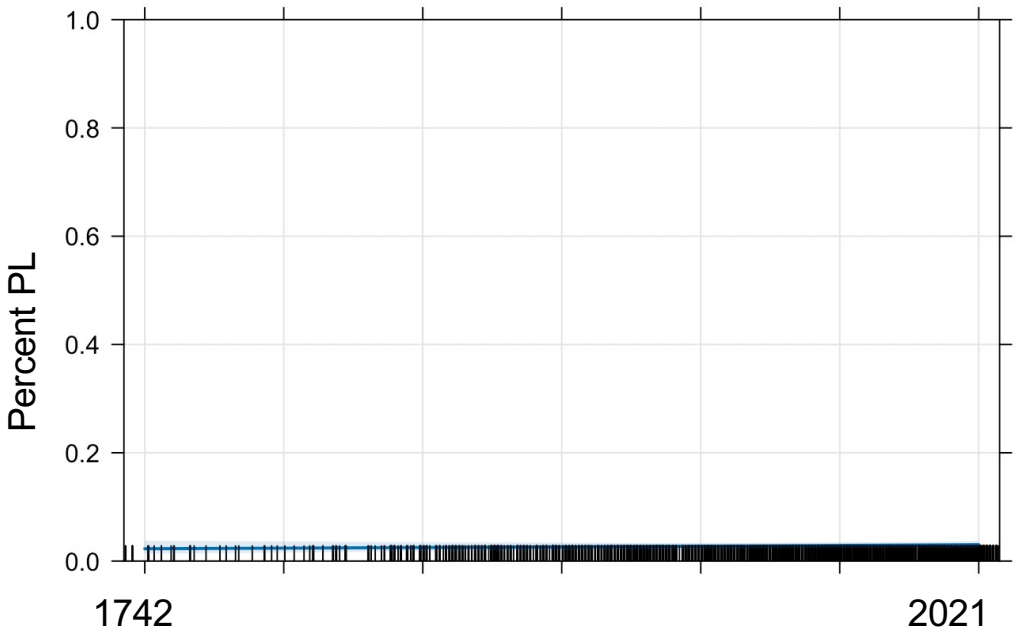
багато людей...
‘many people’

прийшло ‘came’ (SG)

прийшли ‘came’ (PL)

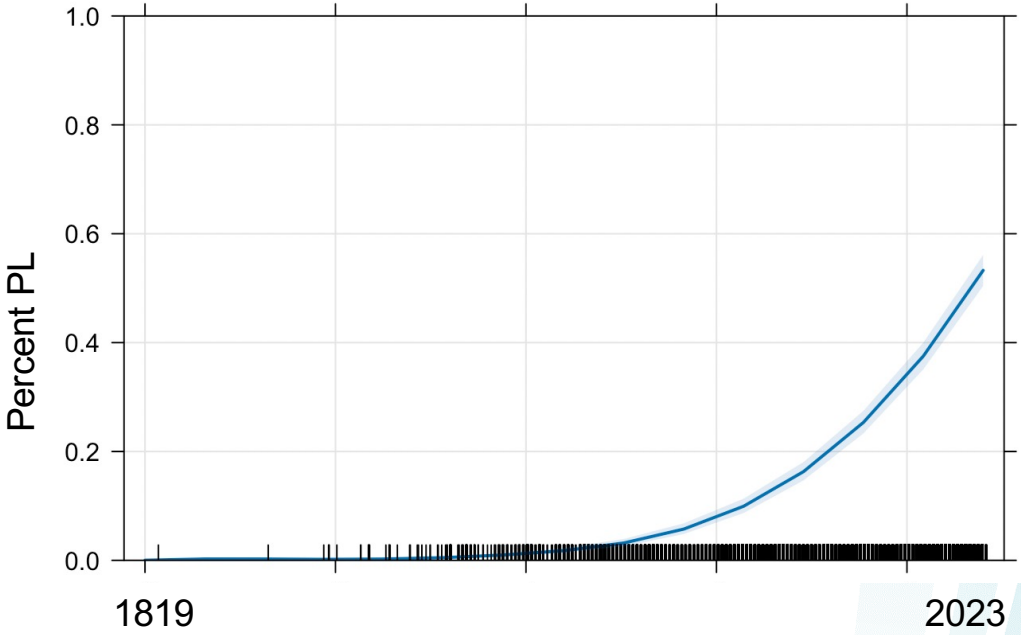
Direction of historical development

Russian *много*



Effect plot based on
logistic regression
6 612 observations from Russian National Corpus
Baseline: 94.7% SG verb forms
(other factors: word order, animacy)

Ukrainian багато

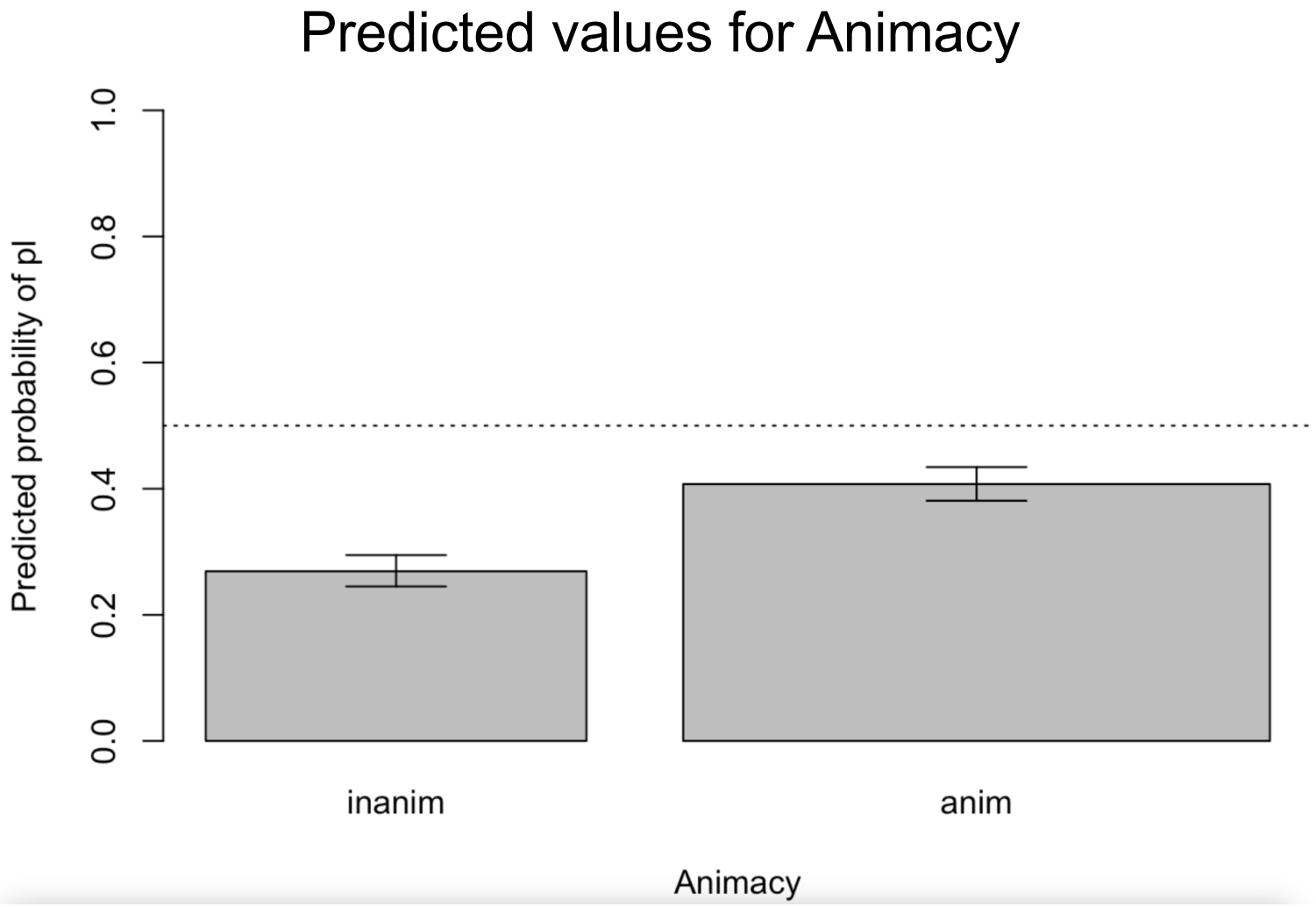


Effect plot based on
logistic regression
28 491 observations from GRAC
Baseline: 69.1% SG verb forms
(other factors: word order, animacy, verb lemma)

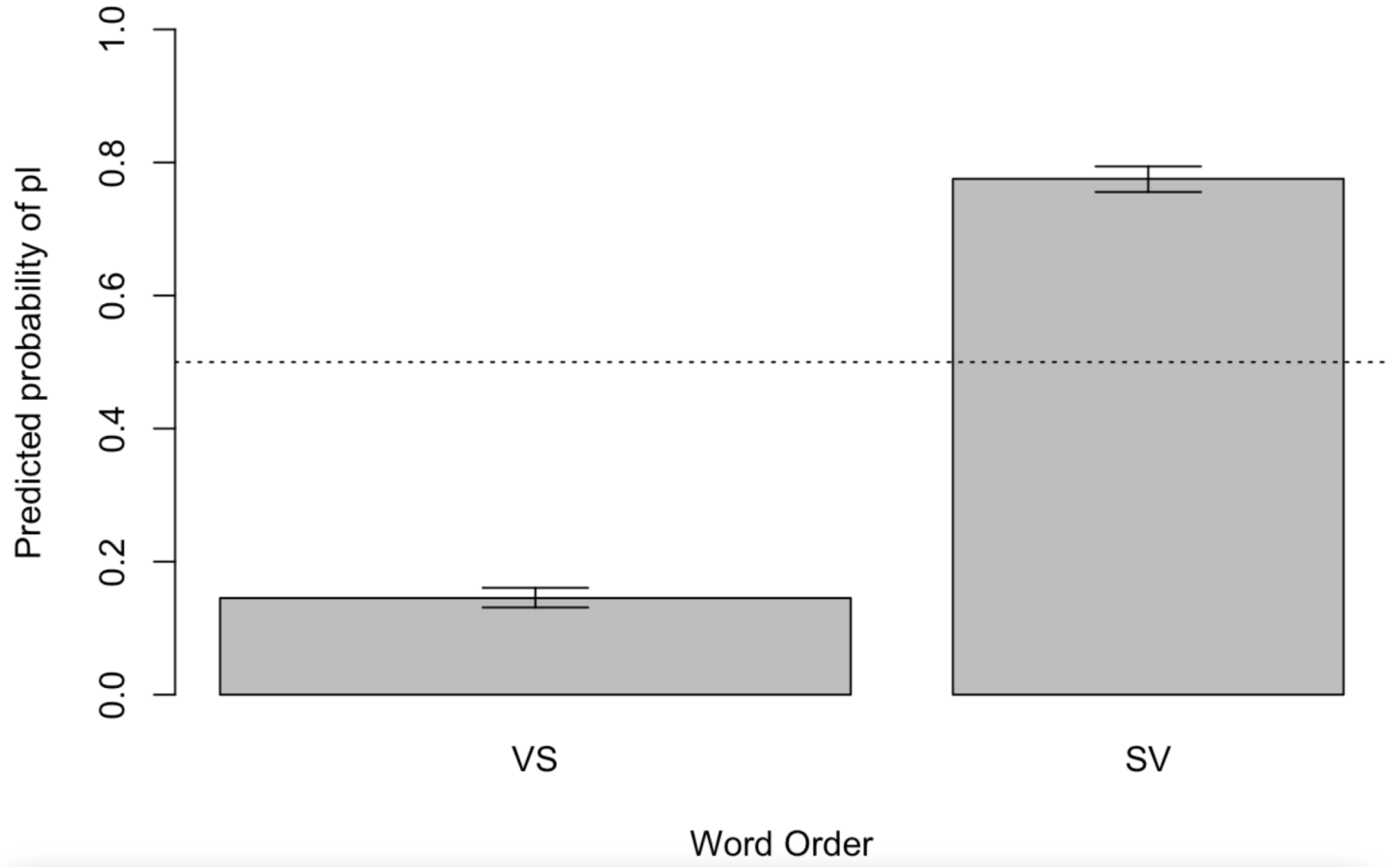
Prediction of Plural by fixed effects of mixed-effect logistic regression model

	Estimate	Standard Error	z value	Pr (>|z|)	
(Intercept)	-2.15067	0.07040	-30.55	< 2e-16	***
Year.sc	0.93924	0.02988	31.44	< 2e-16	***
Animacyanim	0.62370	0.05320	11.72	< 2e-16	***
Word_orderSV	3.01155	0.04991	60.34	< 2e-16	***

Intercept: Year.sc = 0.0 (= 2004), Animacy = inanimate, Word Order = VS


Predicted values for Animacy
Predicted probability of pl
0.0
0.2
0.4
0.6
0.8
1.0
inanim
anim
Animacy

Predicted values for Word Order



Top 10 high-frequency verbs with strongest preference for Singular		Top 10 high-frequency verbs with strongest preference for Plural	
минути ‘pass’	307	*розкритикувати* ‘criticize’	21
відбутися ‘happen’	24	*засудити* ‘condemn’	20
надійти ‘tun up’	204	*ввести* ‘bring in, introduce’	21
виникнути ‘arise’	274	*заявити* ‘declare’	43
виникати ‘arise’	102	*очікувати* ‘expect’	26
накопичитися ‘accumulate’	140	*відзначати* ‘mark, notice’	27
статися ‘happen’	107	*відчути* ‘feel’	30
надходити ‘come’	84	*висловити* ‘express’	50
полягти ‘perish’	44	*отримати* ‘get’	129
назбиратися ‘gather, come together’	52	*пережити* ‘experience’	23

Verb lemmas that appear in twenty or more examples and are ranked by the model at the extremes of a distribution from preference for Singular to preference for Plural. Shading indicates verbs with 100% of observations of only Singular or only Plural.

Case studies of verbs that strongly prefer SG: micro-constructions

- PASSAGE OF TIME
- HAPPEN
- EXIST
- ACCUMULATE
- DIE

Together, the 28 verbs in these 5 semantic groups appear in 12,305 examples, constituting 43% of our total dataset

Group	Verb	N exx	% Sg	% VS	% inanim
PASSAGE OF TIME	минути ‘pass’	307	100 %	87 %	100 %
	пройти ‘pass’	203	78 %	81 %	68 %

І хоча **минуло багато років**, вона про цю подію розповідала з великим хвилюванням. (Інтернет-газета «Високий замок», 2003)
‘And although **many years have passed**, she talked about this event with great excitement.’

Common subjects: *день* ‘day’, *місяць* ‘month’, *рік* ‘year’

Group	Verb	N exx	% Sg	% VS	% inanim
HAPPEN	відбуватися ‘happen’	307	100 %	87 %	100 %
	відбутися ‘happen’	24	100 %	4 %	100 %
	виникнути ‘arise’	274	99 %	89 %	100 %
	виникати ‘arise’	102	99 %	81 %	98 %
	статися ‘happen’	107	99 %	78 %	99 %

Відтоді **багато змін відбулося** в житті актриси і мами. (Інтернет-газета «Дзеркало тижня», 2017)
‘Since then, **many changes have taken place** in the life of the actress and mother.’

Принаймні в автора цих рядків до режисера **виникло багато запитань.** (Інтернет-газета «Україна молода», 2012)
‘The author of these lines at least **had many questions** for the director.’
[literally: there arose many questions]

Group	Verb	N exx	% Sg	% VS	% inanim
EXIST	точитися ‘exist’	46	100 %	91 %	100 %
	існувати ‘exist’	62	100 %	95 %	94 %
	лежати ‘exist’	53	98 %	81 %	62 %
	бути ‘be’	8257	93 %	83 %	54 %
	знайтися ‘exist’	57	100 %	95 %	19 %

when people are referred to, they have limited or subjugated agency

У відділенні **лежало багато хворих** (Онлайн-ЗМІ «UNIAN.NET», 2020)
‘**There were many patients** in the ward’

Group	Verb	N exx	% Sg	% VS	% inanim
ACCUMULATE	нагромадитися ‘accumulate’	23	100 %	83 %	100 %
	накопичитися ‘accumulate’	140	99 %	89 %	100 %
	надійти ‘show up’	204	100 %	90 %	97 %
	надходити ‘show up’	84	100 %	89 %	96 %
	назбиратися ‘gather’	52	100 %	83 %	94 %
	постати ‘appear’	34	97 %	74 %	97 %
	з’явитися ‘appear’	658	98 %	97 %	54 %
	наїхати ‘arrive’	33	100 %	85 %	12 %
	розвестися ‘accumulate’	40	100 %	85 %	10 %
	зібратися ‘gather’	467	95 %	91 %	6 %
	збиратися ‘gather’	66	97 %	83 %	3 %

Відвідувачі Гідропарку скаржаться, що там **розвелося багато щурів**.
(Онлайн-ЗМІ «UNIAN.NET», 2013)
‘Visitors to the Hydropark complain that **there are a lot of rats**.’

Group	Verb	N exx	% Sg	% VS	% inanim
DIE	вимерти ‘die out’	16	94 %	31 %	12 %
	загинути ‘die’	521	82 %	65 %	4 %
	гинути ‘die’	28	82 %	54 %	7 %
	померти ‘die’	96	77 %	28 %	3 %
	полягти ‘perish’	44	95 %	50 %	0 %

І **багато хлопців загинуло** на моїх очах. (Сергій Сущенко, «Спогади», 2018)
‘And **many boys died** before my eyes.’

Summary of micro-constructions

- Verbs associated with five micro-constructions that strongly prefer Singular forms account for 43% of the data in our study
- Most verbs in these micro-constructions prefer VS Word order
- Two micro-constructions are quite distinct: PASSAGE OF TIME, with nouns referring to durations, and DIE, dominated by human beings
- The remaining three micro-constructions are related to each other
- EXIST and ACCUMULATE: inanimate nouns refer primarily to communications and problems, while animates are usually people who are backgrounded in relation to an idea or event

Things that we did not take into consideration

- Distance between the Subject and Verb
- Voice of Verb as Active vs. Passive
- Aspect of Verb as Perfective vs. Imperfective
- Individual author preferences
- Dialectal or regional differences, influence from Russian, Polish
- Other quantifiers
- “Bare” багато without a Noun Phrase (meaning ‘many people’)
- багато з них, багато з нас, багато з вас
- Further semantic classification of nouns and verbs

Ukrainian *bahato*

- In the *bahato* construction the relative frequency in our data is 69.1% Singular vs. 30.9% Plural.
- This relative frequency has changed over time: Plural has increased and recently overtaken Singular, showing a marked difference from Russian.
- SV word order and animate nouns prefer Plural.
- Individual verbs have their own preferences for verb number.
- A handful of micro-constructions that prefer Singular play a large role.

UiT The Arctic University of Norway

Summing up

What will AI do with variation?

- Will AI authentically reproduce variation, representing subtle differences in perception and construal?
- Will AI catalyze change in a stable situation (Russian)?
- Will AI accelerate ongoing change (Ukrainian)?
- Will we even know if there is an impact and what the effects are?