Gender bias in linguistic modelling[1]

Anastasia Gerasimova, Ekaterina Lyutikova
Lomonosov Moscow State University

Dialogue 2022, June 15-18

[1] This research is supported by Russian Science Foundation, RSF project 22-18-00037 realized at Lomonosov Moscow State University, https://rscf.ru/en/project/22-18-00037/ .

1. Gender bias

— The sources for systematic errors in language models.
— Results from representational rates of gendered NPs in professional and social contexts.

Natural language processing (Sun et al. 2019): natural language processing models.
(Stanovsky et al. 2019): both industrial and academic state-of-the-art language models are prone to stereotypical translation errors.

Theoretical linguistics: constructed example sentences.
(Macaulay, Brice 1994, 1997, Cépeda et al. 2021, Kotek et al. 2021): persistent bias toward male-gendered noun phrases (NPs) and perpetuation of stereotypes for both genders.

— Statistical patterns that arise from representational rates of males and females in different professional and social contexts.

— Typicality of certain scenarios is coded both in behavioral manifestations of stereotypes and frequency-dependent parameters of language models.

What is the role of plausibility?

2. Gender imbalance in linguistic examples sentences

Systematic examination of constructed example sentences:

Macaulay, Brice 1994, 1997
Cépeda et al. 2021, Kotek et al. 2021

— A strong tendency to favor male-gendered NPs in comparison to female-gendered NPs.

male-gendered arguments	female-gendered arguments
subjects	nonsubjects
agent and experiencer	patients and recipients
intellectual, perceptual, violent activities	physical, emotional activities
cars, violence, sport, property possession	household chores, romantic relationships
negative emotions	positive emotions

3. Plausibility

— Lack of conflict with knowledge about the world, real-world prototypicality or likelihood of certain argument structure (Tanenhaus et al. 1989, Stowe et al. 1991, Boland et al. 1995, Traxler, Pickering 1996)

— Can ameliorate or lower acceptability regardless of the grammatical status of the sentence (Levelt et al. 1977, Juzek et al. 2018, Schütze 2020)

— Methodology of linguistic experiments: The contribution of implausibility should be minimized in language stimuli.

⇨ Gender bias may result from an unperceived tendency to present acceptability contrasts using the most plausible lexical material.

4. Phenomenon

Criteria in selecting the linguistic phenomenon:

— compatible with various types of situations, with various types of verbs;
— should be polarized, i.e. involve a clear contrast in acceptability;
— the less grammatical stimuli should not be totally unacceptable.

The correlative construction with the correlate *to* 'then'

The polypredicate construction featuring the finite temporal clause headed by the complementizer *kogda* 'when'. [RG80, Inkova 2014, Pekelis 2016, 2018, 2019]

(1) a. ***Kogda*** *Ada* *ušla,* *ded* *zaplakal.*
when Ada.NOM go.PF.PST.F grandpa.NOM cry.PF.PST.M
'When Ada left, grandpa cried.' [RNC, 2014]

b. ***Kogda*** *babuška* *zabila* *v* *konec*
when grandma.NOM hammer.PF.PST.F into tip.ACC
gvozd', ***to*** *paločka* *tresnula.*
nail.ACC then stick.NOM crack.PF.PST.F
'When the grandma hammered the nail into the tip (of the stick), the stick cracked.'
[RNC, 2011]

4. Phenomenon

Different-subject configurations.

The correlative construction involves contrast of the two situations and reinterpretation of the temporal relation between the two situations as the conditional or cause-and-effect relation.

(2) a.*Kogda* *Gerdt* *smejalsja,* *(to)* *v* *mire*
when Gerdt.NOM laugh.IPF.PST.M then in world.LOC
nastupala *garmonija.*
appear.IPF.PST.F harmony.NOM
'When Gerdt was laughing, the harmony appeared in the world.' [RNC 2010]

If such reinterpretation is not supported pragmatically, the different-subject correlative construction is degraded.

(3) *Kogda* *stemnelo,* *(*?to)* *k* *dače* *podkatili*
when get.dark.PF.PST.N then to country_house.DAT roll.PF.PST.PL
dva *gruzovika.*
two truck.GEN
'When it got dark, two trucks rolled up to the country house.' [RNC 2011]

5. Experiment

2x2 factorial design: PRESENCE OF TO STEREOTYPE

(4) a. an agent of aggression predicates (*izbit'* 'beat', *vygnat'* 'expel')
b. an experiencer of sthenic predicates (*obradovat'sja* 'rejoice', *oživit'sja* 'get excited')
c. an agent of predicates denoting professional activities
(*posadit' samolët* 'land an airplane', *razrabotat' model'* 'develop a model')
d. a patient of aggression predicates (*uvolit'* 'fire', *ograbit'* 'rob')
e. an agent of predicates denoting housework
(*pomyt' poly* 'mop the floors', *proteret' pyl'* 'clean the dust')
f. an experiencer of asthenic predicates
(*razrydat'sja* 'burst into tears', *zastesnjat'sja* 'get embarassed')
g. predicates denoting change of state or uncontrolled actions (control)
(*udarit'sja* 'hit oneself', *prostudit'sja* 'catch cold')

(5) Kogda zavod načal vypusk motociklov, …

a. … **Dar'ja** razrabotala novuju model' dvigatelja.	[– to; – stereotypical]
b. … **Dmitrij** razrabotal novuju model' dvigatelja.	[– to; + stereotypical]
c. … **to Dar'ja** razrabotala novuju model' dvigatelja.	[+ to; – stereoptypical]
d. … **to Dmitrij** razrabotal novuju model' dvigatelja.	[+ to; + stereotypical]

'When the factory started producing motorcycles,
Dar'ja / Dmitrij developed a new engine model.'

5. Experiment

- 8 lexical variants for each of the seven predicate types
- 56 sets distributed among 4 lists using a Latin square design
- 56 fillers and four training sentences (½ gram and ½ ungram)
- acceptability judgment task on a 7-point Likert scale
- 64 respondents
- z-transformed ratings
- R environment (lme4, lmerTest)

Model	z score ~ stereotype + to + (1 + to || speaker) + (1 + to || item)			
Fixed effects				
Effect	Estimate	Std.Error	t	p-value
Intercept	0.73669	0.04120	17.883	<< 0.0001 *
stereotypical	0.01558	0.03555	0.438	0.662
to present	-1.24888	0.07700	-16.220	<< 0.0001 *
Random effects				
Grouping	Effect	Variance	sd	Correlation
speaker	Intercept	0.04564	0.2136	
	to present	0.23745	0.4873	-0.96
item	Intercept	0.02028	0.1424	
	to present	0.07178	0.2679	-0.50
Residual		0.35377	0.5948	
Number of obs:	2461	item: 192	speaker: 52	

Figure 2: Estimated fixed and random effects

6. Results

— Responses are significantly lower when *to* is present in the sentence.
— No differences were found with respect to whether the described scenario was stereotypical or not.

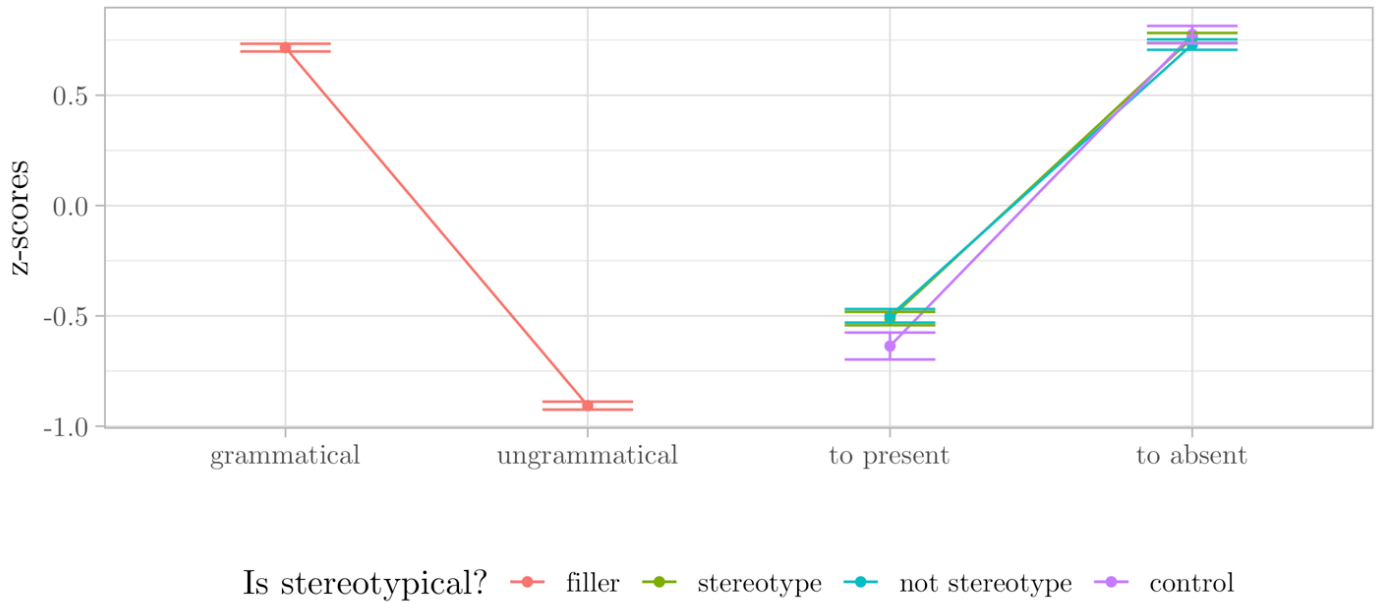

Figure 3: Interaction plot of acceptability ratings (z-score transformed) for target conditions and filler items. Error bars indicate standard error.

7. Conclusion

— No relationship between acceptability and plausibility which can be defined in terms of the stereotypical gender scenarios.

— Eliminating gender bias and stereotyping language from syntactic example sentences does not give rise to implausibility.

— The results of this study do not suggest that the plausibility effect does not exist in principle.

— It is the moral burden of linguists to be attentive to the resulting corpus and to avoid reinforcing stereotypes.

References

Boland J.E., Tanenhaus M.K., Garnsey S.M., Carlson G.N. (1995), Verb argument structure in parsing and interpretation: Evidence from wh-questions, Journal of Memory and Language, Vol. 34, pp. 774–806.

Cépeda P., Kotek H., Pabst K., Syrett K. (2021), Gender bias in linguistics textbooks: Has anything changed since Macaulay & Brice 1997?, Language, Vol. 97, № 4, pp. 678–702.

Inkova O. La corrélation en russe // Structures et interprétations. — Berne: Peter Lang, 2014.

Juzek T.S., Häussler J. Semantic Influences on Syntactic Acceptability Ratings // Proceedings of Linguistic Evidence 2018. — Tübingen: University of Tübingen, 2019.

Kizach J., Nyvad A.M., Christensen K.R. (2013), Structure before Meaning: Sentence Processing, Plausibility, and Subcategorization, PLoS ONE, Vol. 8, № 10.

Kobozeva I. M. Glava 3. Konnektory, vyrazhajushhie otnoshenie neposredstvennogo predshestvovanija odnogo sobytija drugomu [Chapter 3. Connectors expressing relation of immediate precedence of one event to the other] // Struktura konnektorov i metody ee opisanija [Structure of connectors and the means of its description] — Moscow, 2019. — P. 87–117.

Kotek H., Dockum R., Babinski S., Geissler C. (2021), Gender bias and stereotypes in linguistic example sentences, Language, Vol. 97, № 4, pp. 653–677.

Levelt W.J., Van Gent J.A. W.M., Haans A.F.J., Meijers A.J.A. Grammaticality, paraphrase, and imagery. — Acceptability in language, 1977. — P. 87–101.

Macaulay M., Brice C. Gentlemen prefer blondes: A study of gender bias in example sentences // Cultural performances: Proceedings of the Third Berkeley Women and Language Conference. —1994. — P. 449–61.

Macaulay M., Brice C. (1997), Don’t touch my projectile: Gender bias and stereotyping in syntactic examples, Language, Vol. 73, № 4, pp. 798–825.

Pekelis O.E. (2016), Correlative markers, contrastiveness and grammaticalization: A comparative study of conditional correlatives in Russian and Italian, Italian Journal of Linguistics, Vol. 28, №2, pp. 143–180.

Pekelis O.E. (2018), Ellipsis podlezhashchego glavnoi klauzy v russkom yazyke i ego svyaz' s tipologiei korrelyatov [Subject ellipsis in the main clause in Russian and the typology of correlatives], Voprosy Jazykoznanija, №6, pp. 31–59.

Russkaya grammatika: V 2 t. [Russian grammar: in 2 vol.]. — Nauka, Moscow, 1980. — §1672-1681.

Schütze C.T. Acceptability ratings cannot be taken at face value // Linguistic intuitions: Evidence and method. — 2020. — P. 189–214.

Stanovsky G., Smith N.A., Zettlemoyer L. Evaluating gender bias in machine translation. — 2019. — Vol. arXiv:1906.00591. — Access mode: https://arxiv.org/abs/1906.00591

Stowe L.A., Tanenhaus M.K., Carlson G.N. (1991), Filling gaps on-line: Use of lexical and semantic information in sentence processing, Language and speech, Vol. 34, №4, pp. 319–340.

Sun T., Gaut A., Tang S., Huang Y., ElSherief M., Zhao J., Mirza D., Belding E., Chang K.-W., Wang W. Y. Mitigating gender bias in natural language processing: Literature review. — 2019. — Vol. arXiv: 1906.08976. — Access mode: https://arxiv.org/abs/1906.08976

Tanenhaus M.K., Carlson G., Trueswell J.C. (1989), The role of thematic structures in interpretation and parsing, Language and cognitive processes, Vol. 4, №(3/4), SI 211–234.

Traxler M. J., M. J. Pickering. (1996), Plausibility and the Processing of Unbounded Dependencies: An Eye-Tracking Study, Journal of Memory and Language, Vol. 35, pp. 454–475.