Distance to plausible alternatives predicts acceptability ratings in comparative illusion

Introduction: The sentence *More people have been to Russia than I have* is called a *comparative illusion* because it is often judged acceptable while being semantically anomalous[1]. Previous research has (i) identified which linguistic configurations were more illusory (e.g., plural NP than-clause subjects (1b) were more acceptable than singular ones (1a))[2,3] and (ii) provided a noisy-channel account for why a stronger illusion arises with the pronoun than-clause subject (1d-e)[4]. Yet none has explained the variable strength of the illusion effect (i.e., variation in acceptability) across the full range of illusory sentences (1a, 1b, 1d, 1e). This study offers a noisy-channel inspired account: during the acceptability judgment task, comprehenders consider the possibility that the sentence has been corrupted from its intended form. If true, this hypothesis predicts that the edit distance between a perceived illusory sentence and its plausible near neighbor will be anticorrelated with its acceptability. Here we show experimentally that this prediction is correct, contributing to a growing body of evidence that computational models of rational interpretation explain gradience in human processing patterns[5].

Experiment 1: We first replicated prior claims[2-4] that the illusory conditions in (1) vary in acceptability. 500 participants took an acceptability judgment task and each judged the naturalness of 94 sentences (30 critical trials with a within-subjects design in (1), plus 64 plausible fillers) on a 7-point fully labeled Likert scale. A Bayesian multilevel cumulative model (via *brms* in R) supported a steady decrease of acceptability in the NP conditions (plu/cont: β=-0.75, HPD=[-0.86, -0.64], sing/cont: β=-2.3, HPD=[-2.43, -2.20]) whereas the pronoun conditions did not show much of variance (plu/cont: β=-0.25, HPD=[-0.36, -0.14], sing/cont: β=-0.003, HPD=[-0.11, 0.11]).

Experiment 2: To collect the distribution of the plausible alternatives, 200 additional participants made small edits to 30 illusory sentences from Exp. 1 to make them plausible. We applied the Damerau-Levenshtein algorithm[6] to calculate the word-level edit distance between the corrupted sentence and its edited one. A rule-based script coupled with a manual check distinguished 4965 (83.2% of all trials) plausible edited sentences from those that were ungrammatical, unchanged, or drastically different. Fig. 2 shows the mean edit distance by conditions. Common edits involved shifting *more* (e.g., *students have been to Russia more than I have*, dis=2) and forwarding the than-clause with a bare plural NP (e.g., *more students than teachers have been to Russia*, dis=6 from (1a)). A shorter edit distance of a corrupted condition in Fig. 2 maps to a smaller acceptability difference from its corresponding plausible control baseline in Fig. 1.

Statistical analysis: To gauge whether edit distance independently predicts acceptability, we gathered the raw acceptability score of each trial in Exp. 1 on conditions (1a-b) and (1d-e), the log probability of each sentence using GPT-2[7,8], and their mean edit distances in Exp. 2. We first found that sentence log probability and edit distance were not correlated (Pearson's r = -0.018). We confirmed that edit distance uniquely and negatively affects sentence acceptability, with a larger effect size than the log probability (Table 1). In other words, a larger edit distance correlates with a lower acceptability score and a smaller illusion effect for a given illusory sentence.

Conclusion: We identified and experimentally confirmed a novel theoretical prediction from noisy-channel processing theories about the graded strength of comparative illusions.

(1) a. More students have been to Russia than the teacher has. (np, singular, illusory)
b. More students have been to Russia than the teachers have. (np, plural, illusory)
c. More students have been to Russia than teachers have. (np, control, good)
d. More students have been to Russia than I have. (pronoun, singular, illusory)
e. More students have been to Russia than we have. (pronoun, plural, illusory)
f. Many students have been to Russia more than I have. (pronoun, control, good)

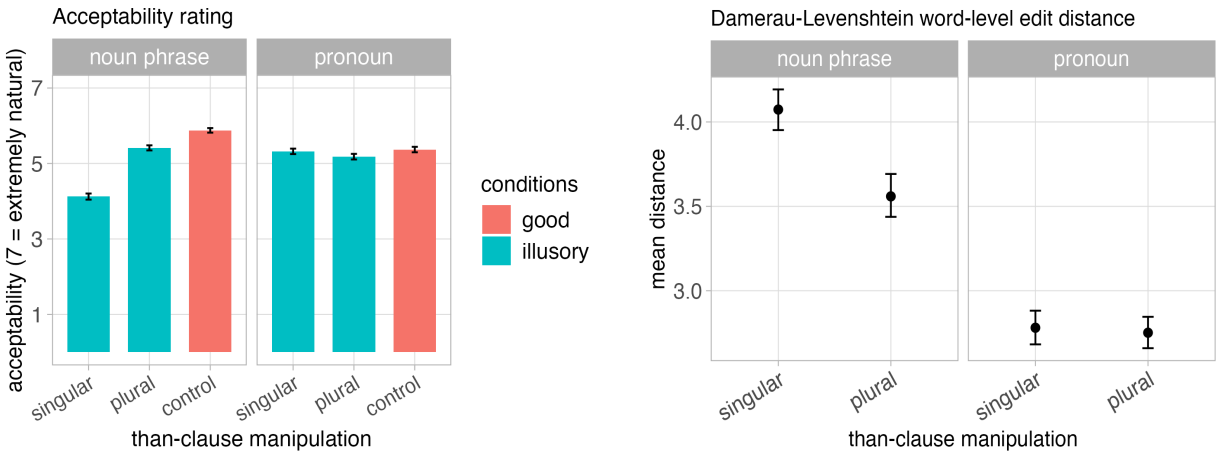

Fig.1 (L) & Fig.2 (R): Error bars are 95% bootstrapped confidence intervals.

Table 1: Statistical modeling results (*lme4* in R)

Model name	Model syntax *y* = raw acceptability rating	Marginal R^2[1]	logProb β (SE), *p*	editDist β (SE), *p*
intercept	~ 1 + (1|participant)	0.0	– –	– –
logProb	~ logProb + (logProb|participant)	0.014	0.23 (0.017), *p* <.001	– –
editDist	~ editDist + (editDist|participant)	0.026	– –	-0.93 (0.068), *p* < .001
Dist_Prob	~ editDist + logProb + (editDist + logProb|participant)	0.037	0.21 (0.016) *p* <.001	-0.89 (0.068), *p* < .001

References:
[1] Montalbetti (1984). PhD thesis. [2] O'Connor (2015). PhD thesis. [3] Wellwood et al. (2018). *J of Semantics*. [4] Zhang et al. (2024). Manuscript. [5] Poliak et al. (2025). Poster at RAILS 2025. [6] pyxDamerauLevenshtein package. [7] Misra (2022). Minicons. [8] Lau et al. (2017). *Cognitive Science*.

[1] A large part of the variance is explained by the random effects. When looking at the conditional R^2 that captures the variance explained by both the fixed and the random effects, it is [0.41, 0.47, 0.43, 0.48]. This does not affect the significance of the fixed effects.