Dense Sentence Sets Induce an Anchor-and-Baseline Strategy in Likert Scale Acceptability Judgments

Artem Novozhilov (artem.novozhilov@ung.si),

Kirill Chuprinko (kirill.chuprinko@ung.si)

Arthur Stepanov (arthur.stepanov@ung.si)
Center for Cognitive Science of Language
University of Nova Gorica
Vipavska 13, SI-5000 Nova Gorica, Slovenia

Abstract

Research in experimental syntax typically assumes that the five-point Likert scale offers an ordinal probe that maps monotonically onto a latent degree of sentence acceptability. We challenge that assumption by showing that, when the stimulus space is densely populated, speakers repurpose the scale into an anchor-and-baseline device. Two large-N experiments (Russian and Serbo-Croatian; N=237; 120 permutational word-order variants per language) elicited over 28000 sentence acceptability ratings. Plotting Shannon entropy of the response distribution against the mean rating reveals a robust 'entropy arch': uncertainty climbs to a sharp peak at the midpoint and collapses toward both ends. We interpret the arch as the quantitative fingerprint of constraint competition: the scale extremes serve as categorical anchors ('completely acceptable' vs. 'completely unacceptable'), while the center functions as a floating baseline against which speakers register maximally uncertain, cue-balanced configurations for which grammatical, information-structural and frequency-based cues pull in opposite directions. Our findings reframe Likert data as the outcome of dynamic calibration rather than static preference strength and provide a simple diagnostic, entropy profiling, for locating linguistic 'tipping-point' constructions. Beyond sentence acceptability, the approach offers a principled way to map regions of maximal competition in any domain where categorical anchors and graded uncertainty coexist.

Keywords: sentence acceptability; Likert scale; entropy, grammaticality

Introduction

Speakers' judgments on syntactic well-formedness remain the main empirical instrument to probe the validity of syntactic theories predicting the grammatical status of sentences. While early theories sometimes viewed grammaticality as a graded notion, the currently established consensus in theoretical and experimental syntax is that grammaticality of a sentence is a theory-internal construct of a binary nature (grammatical, ungrammatical). The gradient character of linguistic judgments is instead better captured by the psycholinguistic notion of acceptability, which reflects not only grammaticality but also various performance factors that influence how a sentence sounds to the speaker. These performance factors include word and n-gram frequencies, plausibility, (implicit) prosody, information structure, richness of morphology, sentence length and perhaps other. Each of these factors can modulate the predicted well-formedness status of the sentence either in the direction of degradation or improvement (Bard & Robertson, 1996; Schütze, 2016; Sprouse, 2007).

Likert scales, often used in experimental syntax, are generally taken to reflect the resulting gradient nature of acceptability by virtue of their monotonic character: a sentence judged with '2' is supposed to sound better than that marked with '1', the one marked with '4' better than one marked with '3' and so on. This is the case with simple 1-to-5 scales as well as more elaborate 1-100 scales used in magnitude estimation studies (Bard & Robertson, 1996; Fukuda, Michel, & Beecher, 2021; Sprouse, 2011). However, the design of stimuli significantly influences the patterns observed in acceptability data. In a typical study in experimental syntax, one compares acceptability of one, or at most a few, sentences of interest against a (minimally different) baseline sentence. Absolute collected values of respective judgments are usually not of primary importance: instead, what is queried is whether there exists a quantitative contrast between the target and baseline mean values, estimated by statistical means (Sprouse, 2007). If the contrast obtains, it is attributed to the factor responsible for the difference between the sentences. This implies a particular *granularity* of the evaluation process which in this case consists of one or few pairwise comparisons, a task quite accessible to even a non-linguistically trained speaker.

It is far less common for studies to compare a large set of sentence stimuli that differ only slightly along a single dimension. In such cases, participants are exposed to a densely packed space of sentence configurations or interpretations, with acceptability judgments spanning the full range of the scale. Given the subtle distinctions between stimuli, speakers are unlikely to engage in extensive pairwise comparisons, as these would be difficult to differentiate. This has immediate implications for the role of baseline reference (whether explicitly provided or not) which becomes significantly diminished. The increased granularity of the task imposes a higher cognitive load compared to tasks with coarser granularity. Consequently, speakers are compelled to rely more heavily on absolute scale values rather than relative comparisons to distinguish between different sentence variants.

This study aims to demonstrate that, when speakers confront a *dense lattice* of minimally contrasting sentences, they spontaneously reorganize the five-point Likert scale into an anchor-and-baseline scheme: the extreme values (1 and 5) act as categorical anchors of certainty, while the midpoint (3) becomes a floating baseline that speakers reserve for constructions where competing grammatical, prosodic and usage-based cues are in equilibrium. By analyzing the joint

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

behavior of mean ratings and Shannon entropy across 120 word-order permutations in Russian and Serbo-Croatian, we show that this strategy produces a characteristic 'entropy arch': low uncertainty at the anchors and a sharp crest at the centre which serves as a quantitative fingerprint of constraint competition.

Experiments

As is well known in the syntactic literature on Slavic languages, speakers exhibit considerable variability in their acceptance of different word order permutations. It is sometimes even claimed that all permutations of a basic SVO word order are acceptable in these languages, though empirical illustrations typically focus on the six permutations of a three-word kernel sentence with canonical word order (Kallestinova, 2007; Stjepanović, 1999). No study to date has systematically examined the full acceptability profile of canonical word order permutations in Slavic languages beyond this minimal sentence length (for four-word kernel permutations in English, see Scott (1969)). Anecdotal evidence suggests that while speakers often have clear intuitions about the acceptability of certain word orders under specific extragrammatical conditions, such as information structure, they frequently hesitate or express uncertainty about others.

Our primary goal in this study was to obtain a statistically robust acceptability profile of the set of sentences that constitute a full set of permutations of the canonical word order in a 5-word kernel, in two free word order languages, Russian and Serbo-Croatian, and to evaluate the effect of a dense stimulus set with increased granularity on the acceptability of individual sentences by measuring the entropy of response patterns per individual word orders as an indicator of the speakers' (un)certainty.

Materials and design

In both experiments, we constructed an exhaustive set of 120 word order permutations of a five-word long kernel sentence with the following structure and canonical word order: [$_{subj}$N] Aux V [$_{obj}$ Adj N]. The subject was always a bare noun phrase, while the object was a noun phrase with an adjective modifier. The auxiliary verb was invariantly *budet* ("will") in Russian, and the singular or plural clitic *je* or *so* indicating a past tense form of 'be' , to ensure grammatical uniformity.

One key issue in designing a study with a dense stimuli set such as ours is deciding on using either (i) the same lexicalization or (ii) different lexicalizations for the entire set. Option (i) might potentially lead to potential carry-over and/or satiation effects (Snyder, 2022), whereas option (ii) when properly balanced will require an enormously large participant pool.

To navigate this trade-off, we conducted two experiments each implementing one of the two respective experimental designs. In Experiment 1 with Russian materials, we employed 120 distinct lexicalizations, carefully selected to prevent participants from relying on morphological cues to predict syntactic role of a word. To achieve this, we systematically varied the subject and object phrases according to animacy, gender, and number. For instance, if the subject was animate, feminine, and singular, the object was inanimate, masculine, and plural, and vice versa. This approach helped mitigate potential morphologial cues for participants. The lexicalizations were distributed across six experimental lists. Each list contained 20 non-canonical word orders spanning six distinct lexicalizations, resulting in a total of 120 sentences per protocol, representing 20 construction types. The assignment of constructions to protocols was randomized to prevent order effects.

Experiment 2 with Russian materials employed a simpler design with only a single lexicalization for each protocol. In the Serbo-Croatian version of this experiment, three different protocols each involving a different lexicalization were implemented. Representative kernel sentences in Russian and Serbo-Croatian are in (1) and (2), respectively. All materials for this experiment, inclduding the instructions as well as the results, are available on OSF.

(1) *Mekhanik* *budet* *chinit'*
Mekhanik.NOM.SG.M. budet chinit'.INF.IMPF
gruzovye *mashiny.*
gruzovye.ACC.PL. mashiny.ACC.PL.F
'The mechanic will repair cargo trucks.'

(2) *Radnici* *su* *izgradili*
workers.NOM.PL. BE.PL. built.PART.
novu *zgradu.*
new.ACC.SG.F. building.ACC.SG.F.
''The workers have built a new building.'

Participants

Overall, two hundred thirty seven adult native speakers of Russian and Serbo-Croatian took part in this study. 79 adult self-reported native Russian speakers took part in Experiment 1 (52 female, 2 other; mean age = 30.1). 40 native Russian participants (28 female; mean age = 28.5) who did not participate in Experiment 1, took part in Experiment 2 with Russian materials. One hundred eighteen adult self-reported native speakers of Serbo-Croatian participated in the Serbo-Croatian version of Experiment 2 (85 female, median age=33) and were assigned to one of the three single lexicalization protocols on a random basis. None of the participants reported any history of neurological conditions. While 19 participants from the Russian group disclosed previous mental health issues, neither their acceptability ratings, nor response time did not differ significantly from those of the rest of the population. Participants accessed the experiment via the PCIbex platform (Zehr & Schwarz, 2018), which was programmed to automatically assign respective groups of participants randomly across protocols.

Procedure

Participants rated acceptability of the stimuli sentences on a 5-point Likert scale, where the extreme points were marked as 'completely unacceptable' and 'completely acceptable' (in the respective language); the remaining scale values were not qualitatively marked. The task followed a speeded paradigm, allowing a maximum of 7 seconds per sentence. Participants were asked to follow their first intuition in evaluating the sentences and not to dwell on their judgment. To mitigate fatigue, participants could take short breaks after every 20 sentences. Sentences were presented individually, and participation was restricted to PCs and laptops.

Data Analysis and Results

For data analysis, we used R (R Core Team, 2021) and conducted statistical modeling with linear mixed-effects models using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). This approach allowed us to account for both participant variability and presentation order as random effects, ensuring that individual differences and sequence effects did not bias the results. To perform multiple comparisons, we used the multcomp package (Hothorn, Bretz, & Westfall, 2008), applying Tukey-adjusted post-hoc tests to assess pairwise differences between methodological conditions. This allowed us to determine whether the choice of lexicalization strategy influenced acceptability ratings.

The results, summarized in Table 1, compare the three methodological conditions: single lexicalization methodology (1), three-lexicalizations methodology (2), and multiple lexicalizations methodology (3). As indicated by the multiple comparisons, there were no statistically significant differences between these conditions.

Table 1: Pairwise Tukey-adjusted contrasts among methodological conditions in the mixed-effects linear regression predicting rating.

Comparison	Estimate	Std. Error	z value	Pr(> |z|)
$2-1=0$	-0.2096	0.6611	-0.317	0.945
$3-1=0$	1.0295	0.6666	1.544	0.267
$3-2=0$	1.2390	0.8133	1.523	0.276

Moving on to the entropy analysis, individual acceptability profiles for each tested word order were compiled based on each experiment and language, with the percentage of occurrence of each rating plotted (Figure 1).

Based on this distribution, respective probabilities of occurrence for each rating and entropy per word order were calculated. The notion of entropy, as formulated in information theory by Shannon (1948), was used as a measure to quantify the level of unpredictability or uncertainty within a system. Mathematically, entropy H(x) is defined as:

$$H(X) = -\sum_{i=1}^{N} p_i \log_2 p_i, \quad (1)$$

where p_i represents the probability of each possible outcome i in a given set of N possible states.

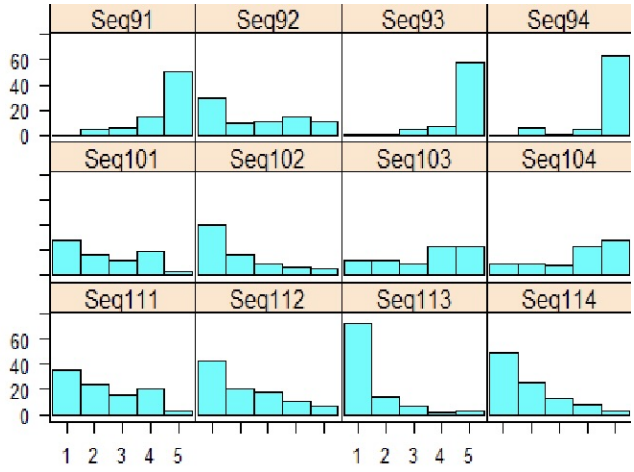

Figure 1: Experiment 2, Serbo-Croatian stimuli: A segment of the set of Likert rating distributions per word order permutation (marked 'Seq'), in percent of occurrence.

If the ratings tend to cluster around a particular value (e.g. 1 or 5), this indicates a lesser entropy or uncertainty in delivering the judgment. In contrast, a larger spread across the scale indicates a greater uncertainty, signaled by heightened entropy. The maximum entropy would correspond to a uniform distribution where each value would be represented approximately equally across the scale (cf. word orders Seq103 and Seq104 on Figure 1). The maximum entropy on a 1-to-5 Likert scale can be calculated thus (assuming all responses are a priori equally probable).

$$H(X) = -5 \times \frac{1}{5}\log_2\frac{1}{5} = -\log_2\frac{1}{5} = \log_2 5 \approx 2.32 \text{ bits} \quad (2)$$

Using this measure, we examined the relationship between entropy and mean acceptability ratings across all 120 word orders. This is plotted for Experiment 1 in Russian in Figure 2 and Experiment 2 in Serbo-Croatian in Figure 3.

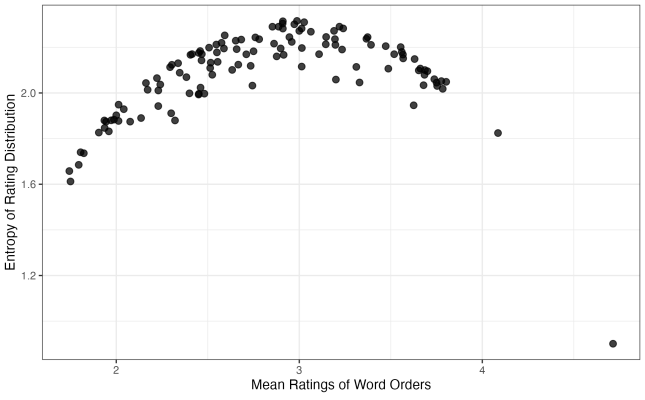

Figure 2: Relationship between entropy and mean acceptability ratings in Experiment 1 (Russian, multiple lexicalizations)

Both distributions of entropy follow a parabolic pattern. The highest entropy values are associated with word order

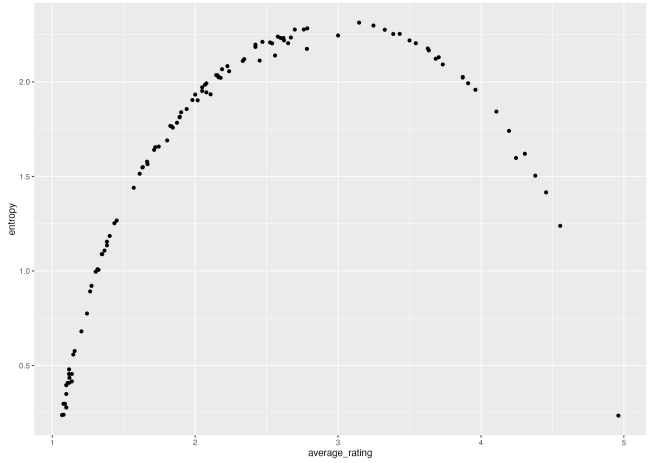

Figure 3: Relationship between entropy and mean acceptability ratings in Experiment 2 (Serbo-Croatian, average of three single lexicalizations)

combinations that received a mean acceptability rating of or around 3, indicating maximal uncertainty in judgments. These mean ratings correspond to (near-) uniform distributions of ratings according to the Cramer-von Mises test of goodness-of-fit (p>0.05). The lowest entropy value corresponds to the word orders rated at the extreme points of the scale (1 and 5), where ratings exhibit minimal variability, suggesting, instead, strong agreement among participants. To further illustrate this point, Table 2 provides a comparative illustration between probability distributions for each rating along with the points of highest entropy and corresponding average rating (here we provide only the 15 representative word order in Experiment 1 (Russian) to keep the article concise. The full table can be found at OSF. For brevity, we abbreviated specific word positions: [$_{subj}$N] = 1, Aux = 2, V=3, [$_{obj}$Adj] = 4, [$_{obj}$N] = 5.

Because Shannon entropy is a strictly concave function of the full response distribution, any variable with fixed, finite categories will trace an inverted-U 'arch' when entropy is plotted against the mean rating. That shape itself is therefore not diagnostic. What is informative is where our stimuli land on that surface. In our data, sentences that combine harmonious grammatical, information-structure and frequency cues cluster at the low-entropy flanks (means $\approx$ 1.3 and 4.7), whereas constructions in which those cues pull against each other rise to the high-entropy crest (mean $\approx$ 3.0). Thus, the distribution of points over the arch captures the equilibrium (or lack thereof) among competing constraints, giving us a concrete linguistic diagnostic for identifying sites of strongest constraint interaction (see below).

Discussion

A typical Likert scale used in acceptability studies can be considered a unipolar scale. Unipolar scales have two defining properties: (i) they measure the degree of presence of a particular property, and (ii) they lack a natural midpoint when approaching a single extreme. The highest uncertainty observed at the mid-scale, as a task effect, suggests that speakers performing an acceptability task with a dense stimulus set do not primarily use the mid-scale value (3) to indicate a gradient of acceptability. Instead, the mid-scale value functions as an aggregate mean of individual judgments distributed across the entire scale. In other words, rather than relying on mid-range values, speakers tend to use values closer to both extremes. This suggests that, in such cases, speakers interpret the Likert scale more as an emergent bipolar or bipolar reversal scale anchored at the extreme points. Unlike unipolar scales, bipolar scales (i) measure evaluation between two opposing extremes and (ii) assume a natural midpoint (Shulman, 1973; Wang & Krosnick, 2020).

What does this mid-scale interval correspond to in the speaker's mental evaluation process? As noted earlier, an acceptability judgment is a function of both grammaticality and various performance constraints. Across different word orders, these constraints carry different 'weights' in the speaker's evaluation mechanism, influencing judgments toward either extreme of the Likert scale. Notably, when performance constraints conflict with grammatical judgment, they may not entirely override it but instead shift the perceived acceptability of an ungrammatical sentence toward a higher rating. A well-known example of this phenomenon is the comparative illusion as in *More people have been to Russia than I have*, where structural and semantic factors create an impression of acceptability despite an underlying grammatical violation (Saddy & Uriagereka, 2004).

We hypothesize that the word orders with the highest observed entropy are those where the weights of all the factors or constraints involved in the speaker's evaluation are balanced. In these cases, all constraints have similar or equal weights, and none tips the scale in either direction. The resulting ambiguity leads to a kind of response that becomes largely a matter of individual speaker variation, leading to a uniform distribution of ratings. Consequently, the mid-range interval can be seen as an abstract 'standard' or 'baseline' reference, against which speakers evaluate other word orders, moving in either a positive or negative direction depending on which constraints outweigh the others. This standard reference point effectively splits the scale in half suggesting a binary aspect of the evaluation process. Due to the high density of evaluation items in the permutation set, this binary process becomes apparent as speakers notably avoid the midpoint scale value for a subset of stimuli.

This perspective also clarifies the role of the scale midpoint. The central rating does not mark 'moderate acceptability' in any absolute sense; rather, it functions as a floating baseline that speakers reserve for exactly those configurations where various linguistic and extra-linguistic constraints cancel each other out. By contrast, the scale extremes serve as categorical anchors of certainty: 'completely unacceptable' at one end, 'completely acceptable' at the other.

Because the entropy-mean profile pinpoints sentences in equilibrium, it offers a practical diagnostic for locating linguistic 'tipping-points' worthy of closer experimental or theo-

[h]

Table 2: Entropy values and probability distributions for different word order constructions.

Combination	Entropy	$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$	*Av.rating*
5 1 3 4 2	2.32	0.18	0.23	0.22	0.19	0.19	2.99
2 3 4 5 1	2.31	0.22	0.19	0.23	0.18	0.18	2.91
3 1 4 5 2	2.31	0.16	0.23	0.23	0.20	0.19	3.03
5 3 1 2 4	2.30	0.21	0.23	0.21	0.14	0.21	2.91
3 2 4 5 1	2.30	0.22	0.18	0.24	0.15	0.22	2.97
4 1 2 3 5	2.29	0.17	0.15	0.20	0.27	0.22	3.22
4 5 3 2 1	2.29	0.26	0.16	0.23	0.15	0.21	2.89
2 5 1 3 4	2.29	0.23	0.19	0.27	0.15	0.17	2.85
1 3 5 4 2	2.28	0.12	0.19	0.24	0.24	0.21	3.24
1 4 3 2 5	2.28	0.15	0.22	0.28	0.15	0.19	3.01
3 5 1 2 4	2.28	0.22	0.18	0.21	0.27	0.13	2.91
1 5 3 2 4	2.27	0.13	0.18	0.25	0.27	0.18	3.19
5 2 4 1 3	2.27	0.13	0.26	0.26	0.20	0.16	3.00
4 1 5 3 2	2.27	0.14	0.20	0.30	0.16	0.19	3.06
4 3 2 5 1	2.25	0.28	0.22	0.23	0.14	0.12	2.59

retical scrutiny. Many other factors including processing ease or pragmatic felicity may also shape the entropy landscape. Future work should enlarge the cue inventory, explore languages with different word-order flexibility, and vary lattice density to test how robust the anchor-and-baseline strategy is across tasks and populations.

Other relevant studies

Several recent studies suggest that the effect of the highest variance in the mid-zone is not artifactual to word order variations and extends to other types of dense stimuli sets. Brown, Fanselow, Hall, and Kliegl (2021) ran an acceptability judgment task on a set of German and English sentences that were pre-calibrated (= pre-tested in different with other speakers) across the entire spectrum of acceptability ranging from fully unacceptable to fully acceptable. They found that the syntactic satiation effect (improvement of the speaker's subjective judgment regarding a sentence type over a repeated exposure to it) arises irrespective of sentence type, for those sentences whose acceptability status falls in the mid-zone range of a discrete Likert scale. In a follow up study, Stepanov (2024) demonstrated that mid-scale ratings in that study form a region of highest uncertainty as reflected in maximum variance in speakers' ratings compared to the other regions of the scale. Satiation may consequently be seen as an exposure effect targeting the most unstable or 'volatile' portion of the judgments. In our present terms, if the variance in ratings is indeed due to speakers' subjective uncertainty about the 'unacceptable' and 'acceptable' halves of the scale, then satiating at least in some cases means switching simply from the former to the latter half, which fits in the context of the binary character of the speakers' evaluation process (Schütze, 2016).

There is also some indication that the anchor-and-baseline character of speaker's evaluation in dense stimuli set extends beyond syntactic acceptability tasks, to interpretational aspects of the sentence evaluation. Stateva, Stepanov, Déprez, Dupuy, and Reboul (2019) probed approximate numerical boundaries associated with inherently vague quantifiers such as *some* in English, German, French and Slovene. The task was to evaluate how well sentences with a given quantifier describe a situation where the actual number of the quantified individuals is given in relation to the total number of relevant individuals in a given context. For instance, the subjects were asked to evaluate a sentence like *Some men utilized an online dating site* in contexts such as *133 men sought a life partner. 41 of these men utilized an online dating site*, on a 1-to-5 Likert scale, in a speeded manner (no actual arithmetic calculations were allowed). The researchers manipulated the proportion of the quantified expression by varying the second number in the supporting context so as the resulting ratio would be between 1-99% of the total with an increment of 2% resulting in a dense stimulus set yielding 50 data points per quantifier per subject. Notably, dispersion of judgments measured by standard deviation in that study turned out to vary practically as a mirror image of the rating pattern increasing toward the middle (3 out of 5) and decreasing toward the scale extremes, forming the familiar parabolic pattern overall. This suggests that the abstract 'standard' of evaluation was entrenched in that evaluation process for a certain subset of contexts as well and speakers evaluated the rest of the contexts in relation to that latent standard, following a binary, rather than gradient, pattern.

Concluding remarks

It is tempting to relate the observed anchor-and-baseline evaluation strategy to the binary nature of a sentence's grammaticality status associated with both positive and negative extremes (see Introduction and Schütze (2016)). However, the relationship is not immediately clear: as mentioned earlier, grammaticality is a theory-internal construct, whereas the anchoring classification is a bias speakers adopt dynamically when faced with a dense stimuli set. Additionally, as noted above, the measured acceptability ratings reflect non-grammatical performance factors, which may play a significant role in influencing the final judgment, though it is not

clear whether their impact is less significant than grammaticality. A more precise mapping between the graded and binary components of the speakers' sentence evaluation mechanism is needed, and this remains a subject for future research (see, e.g. Bader and Häussler (2010), for a relevant discussion).

An alternative explanation for the anchor-and-baseline pattern may lie in cognitive economy. When the full set of linguistic cues such as grammar, information structure, frequency, prosody etc. converge on the same verdict, the processor can reach a decision quickly and with high confidence, producing low-entropy, extreme ratings. By contrast, at the scale midpoint the same cues pull in opposing directions, forcing the system to integrate conflicting evidence; that additional deliberation manifests as higher entropy and longer reaction times. Future work should test this load-sensitive model directly, for instance by correlating entropy with response latencies or neural markers of cognitive effort.

To summarize, our findings suggest that gradient sentence-acceptability judgments emerge from the interaction of multiple linguistic and extra-linguistic cues whose relative strengths vary across constructions. In densely sampled stimulus spaces, speakers push the extremes of the Likert scale toward categorical anchors for sentences where these cues converge, while the midpoint becomes a floating baseline for cases where the cues balance one another, producing maximal uncertainty. This anchor-and-baseline strategy becomes more pronounced as stimulus granularity increases, a pattern echoed in other dense-design studies. Future work should extend entropy profiling cross-linguistically and combine detailed cue-weight models with formal syntactic analysis to clarify how competing constraints are integrated during acceptability decision-making.

Acknowledgments

This research has received funding from the Slovenian Research and Innovation Agency (ARIS) under project no. J6-4615.

References

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, *46*(2), 273–330.

Bard, E., & Robertson, D. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32–68.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brown, J. M. M., Fanselow, G., Hall, R., & Kliegl, R. (2021). Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *PLoS One*, *16*(5), e0251280.

Fukuda, S., Michel, D., & Beecher, H. (2021). Is magnitude estimation worth the trouble? In G. Goodall (Ed.), *Theory and experiment in syntax.* New York, NY: Routledge.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Kallestinova, E. D. (2007). *Aspects of word order in Russian*. Doctoral dissertation, University of Iowa, Iowa City.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `https://www.R-project.org/`

Saddy, D., & Uriagereka, J. (2004). Measuring language. *International Journal of Bifurcation and Chaos*, *14*(2), 383–404.

Schütze, C. T. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.

Scott, R. I. (1969). A permutational test of grammaticality. *Lingua*, *24*, 11–18.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shulman, A. (1973). A comparison of two scales on extremity response bias. *Public Opinion Quarterly*, *37*(3), 407.

Snyder, W. (2022). On the nature of syntactic satiation. *Languages*, *7*(1), 38.

Sprouse, J. (2007). *A program for experimental syntax*. Doctoral dissertation, University of Maryland, College Park, MD.

Sprouse, J. (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, *87*(2), 274–288.

Stateva, P., Stepanov, A., Déprez, V., Dupuy, L. E., & Reboul, A. C. (2019). Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, *10*, 957.

Stepanov, A. (2024). Satiation and uncertainty in the midzone of sentence acceptability judgments. *Linguistische Arbeitsberichte*, *96*, 323–335.

Stjepanović, S. (1999). *What do second-position cliticization, scrambling and multiple wh-fronting have in common?* Doctoral dissertation, University of Connecticut, Storrs.

Wang, R., & Krosnick, J. A. (2020). Middle alternatives and measurement validity: A recommendation for survey researchers. *International Journal of Social Research Methodology*, *23*(2), 169–184.

Zehr, J., & Schwarz, F. (2018). PennController for internet-based experiments (IBEX).
doi: 10.17605/OSF.IO/MD832