Distance and Projectivity as Predictors of Sentence Acceptability in Free Word Order Languages

Kirill Chuprinko
Center for Cognitive Science of Language
University of Nova Gorica
Vipavska 13, 5000 Slovenia
kirill.chuprinko@ung.si

Artem Novozhilov
Center for Cognitive Science of Language
University of Nova Gorica
Vipavska 13, 5000 Slovenia
artem.novozhilov@ung.si

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

Abstract

This study investigates how two core metrics rooted in Dependency Grammar, Mean Dependency Distance (MDD) and projectivity, predict sentence acceptability in Russian and Serbo-Croatian. Using exhaustive word order permutations in controlled five-word sentences, we model how these metrics relate to acceptability judgments in two psycholinguistic experiments. While MDD has been widely studied as a processing constraint, projectivity violations have received less attention in experiments, and particularly in acceptability modeling. We demonstrate that both metrics have a significant independent impact on judgments, with projectivity playing a surprisingly strong role. In addition, Serbo-Croatian's rigid clitic placement provides a natural test case for disentangling grammatical from processing constraints. Our findings offer a computationally precise, dependency-based model of acceptability that advances cognitively grounded language modeling for free word order languages.

1 Introduction

Sentence acceptability reflects how natural or well-formed a sentence appears to native speakers, bridging linguistic competence and real-world performance (Chomsky, 1965). While judgments of acceptability are shaped by multiple factors such as semantic plausibility, discourse coherence, and real (for listening mode) or potential prosody, they are fundamentally influenced by two key forces: i) grammaticality (conformity to internalized rules of grammar) and ii) processing load (constraints arising during sentence comprehension and production). Cognitively informed models of sentence acceptability must capture both of these dimensions. Yet much of the current research, particularly in the evaluation of neural language models, tends to conflate them or focus on surface-level performance metrics (Warstadt et al., 2019; Zhang et al., 2024). This limits our understanding of the underlying mechanisms that drive acceptability.

One well-established processing constraint is *dependency distance*, rooted in the framework of Dependency Grammar (Mel'čuk, 2009). Prior work based on Universal Dependencies (Futrell et al., 2015; Choi, 2007; Ros et al., 2015) demonstrated a general tendency for speakers for shorter syntactic dependencies across languages. This is formalized in the "Minimize Dependency Distance" principle (MDDP) and operationalized in Mean Dependency Distance (MDD) as a metric. MDD quantifies how far apart syntactically related words appear in linear order. Increased distance is thought to increase processing cost and, by extension, reduce acceptability. This naturally aligns with memory-based theories of sentence processing (Gibson, 1998, 2000).

However, several questions remain open. First, how does MDD interact with grammatical constraints, especially in languages with relatively free word order? Second, is MDD the only processing-related factor influencing acceptability, or are other surface structural properties, such as projectivity violations, where dependency arcs cross (Testelets, 2001; Gildea and Temperley, 2010; Liu et al., 2017; Yadav et al., 2020, 2022)? While MDD has been well integrated into psycholinguistic models of sentence processing, projectivity remains largely underexplored in experimental linguistics, especially in the context of acceptability judgments.

This study addresses these gaps by systematically modeling sentence acceptability across all possible word order permutations in five-word sentences in Russian and Serbo-Croatian. The choice of these two languages is not accidental. Both languages allow documented high word order flexibility, with some theoretical studies claiming that any permutation of words in a clause is acceptable (Kallestinova, 2007; Stjepanović, 1999). At the same time, only Serbo-Croatian poses a hard grammatical constraint on word order: clitics must be in the second position. In both languages, permuting words in a sentence leads to variation in syn-

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 108–119

tactic dependency lengths which can be captured by MDD but also be gauged by other word-order related metrics. By systematically varying word order while controlling for lexical and structural factors, the role of different dependency metrics can be isolated. Furthermore, a comparison of sentence acceptability profiles in these two otherwise close languages allows us to disentangle the grammatical factor from processing effects.

We evaluate several dependency-based metrics including MDD, projectivity violations and a number of other structural and processing complexity measures, on their ability to predict sentence acceptability. Our results show that both MDD and projectivity violations significantly contribute to acceptability ratings, with projectivity playing a stronger role than previously assumed. These findings offer a computationally precise, cognitively grounded model of acceptability that applies core principles of Dependency Grammar to a psycholinguistic context, while deepening our understanding of word order preferences in free word order languages.

2 Minimize Dependency Distance Principle and Its Measures

The MDDP is rooted in cognitive constraints associated with working memory and retrieval processes during sentence comprehension and production. The key idea is that syntactically related words should be placed closer together to facilitate efficient processing and reduce cognitive load.

The first intuitions suggesting the existence of a cognitive mechanism responsible for favoring shorter dependencies in sentence structure can be traced back to the early 20th century. In his descriptive study, Behaghel (1909) noted that in German ditransitive constructions, longer noun phrases typically follow shorter ones. This observation can retrospectively be attributed to the MDDP (Staub et al., 2006).

The second major step toward the formulation of the principle came from psycholinguistic research in the 1970s and 1980s, when psychologists (Perfetti and Lesgold, 1977; Daneman and Carpenter, 1980) showed that comprehension difficulty is affected by the amount of information that needs to be actively maintained and retrieved, depending on working memory capacity. These findings provided the empirical foundation upon which the principle was explicitly formulated in the 1990s in a range of grammatical and cognitive approaches (Hawkins, 1994; Hudson, 1995; Gibson, 1998).

Since its formal articulation in the 1990s, the principle has received substantial empirical and theoretical support from studies on sentence processing, working memory, and syntactic dependency structures (Gibson, 2000; Ferrer-i Cancho, 2004; Choi, 2007; Liu, 2008; Futrell et al., 2015). One of the most influential formulations is Gibson's Dependency Locality Theory (DLT) (Gibson, 2000), which posits that processing difficulty increases with the linear distance between syntactically dependent elements, as longer dependencies demand more memory resources to maintain. This perspective is further elaborated by retrieval-based approaches to parsing, which argue that sentence processing involves cue-based retrieval from memory. According to Lewis and Vasishth (2005), greater syntactic distance increases the likelihood of interference and retrieval failure, thereby raising processing cost. Minimizing dependency length, therefore, enhances the accessibility of syntactically related words, reduces interference, and facilitates more efficient parsing (Grodner and Gibson, 2005; Lewis and Vasishth, 2005).

While the conceptual foundation of the MDDP is widely accepted, its mathematical operationalization varies across studies. Researchers propose different ways to quantify dependency length. For example, Gibson (2000), working in the context of phrase structure grammar, computes the incremental integration cost of each dependency as the number of intervening discourse-referent words and adds this to the concurrent storage cost (i.e. the number of yet-to-be-resolved dependencies being held in memory). Within the dependency grammar tradition, it is more common to use the mean dependency distance as proposed by Liu (2008), which averages the absolute linear distances between heads and dependents. A related, though less widespread, approach comes from Ferrer-i Cancho (2004), who suggests using the mean Euclidean distance, defined as the average of $\sqrt{(P(\text{head})_i - P(\text{dep})_i)^2}$ – a method more common in computational linguistics (Futrell et al., 2020). Although Ferrer-i-Cancho's approach involves squaring and taking the square root of the difference, in one-dimensional space (i.e., linear word order), it produces the same results as Liu's simpler absolute-value method.

For this reason, in this work we adopt a uniform and computationally straightforward definition of

MDD (see above) along the lines of Liu (2008), calculated as follows:

$$\text{MDD} = \frac{1}{n-1} \sum_{i=1}^{n} |P(\text{head})_i - P(\text{dep})_i| \quad (1)$$

where n is the number of words in the sentence and P denotes the position of a word in the linear sequence.

3 Russian and Serbo-Croatian Word Order

Word order in Slavic languages is often described as relatively free, with variations influenced by both grammatical constraints and processing considerations. The canonical (i.e., most frequent or "default") word order in both Serbo-Croatian and Russian is Subject–Verb–Object (SVO) (Urošević et al., 1986; Bailyn, 1995). Adjectives typically precede nouns in both languages. Additionally, while Russian employs almost no tense auxiliaries except for the Future Imperfective form of 'to be,' Serbo-Croatian extensively uses the auxiliary 'to be' to form both past and future tenses. In both languages, auxiliaries follow the subject and precede the verb. A key difference, however, is that in Serbo-Croatian, tense auxiliaries are clitics obeying Wackernagel's Law, meaning they must always appear in the second position in the sentence (Bošković, 2001). However, the canonical sentence structure in both languages is the same: Subj Aux Verb [Adj Obj] (Bailyn, 1995; Bošković, 2001; Bošković, 2005).

How does the MDD fit into this picture? Suppose we define syntactic dependencies as follows: Verb–Auxiliary, Verb–Subject, Verb–Object, Object–Adjective (de Marneffe et al., 2014), or alternatively, Auxiliary–Verb, Auxiliary–Subject, Verb–Object, Object–Adjective (Groß and Osborne, 2015). From the perspective of processing cost (see above), a reasonable hypothesis is that, for a given sentence, the greater the MDD, the lower its acceptability rating. However, this reasoning does not take into account independent grammatical constraints on word order. Due to the strict second-position requirement for clitics, any deviation from Wackernagel's Law in Serbo-Croatian (but not in Russian) would cause a downgrade in acceptability, regardless of the MDD. Because the categorical second-position rule and the gradient, dependency-based memory pressures captured by MDD pull in different directions in Serbo-Croatian, but not in Russian, the two languages jointly provide an ideal testbed for disentangling how rigid grammatical constraints and processing costs on syntactic dependencies shape word-order acceptability.

4 Other Word-Order Related Metrics

As pointed out above, the MDD is not the only metric that can account for patterns in acceptability judgments in the so-called "free word order" languages. The literature proposes several additional ways to quantify how much a sentence deviates from its canonical order, yet none has been systematically evaluated in an experimental setting. In this work we explore five additional metrics: *Number of Displaced Words*, *Total Path of Displaced Words*, *Number of Projectivity Violations*, *Word Order Penalty Score* and *Special Status Residuals and Processing Penalties*. These metrics were developed and/or adapted for the present study drawing on insights from a wide range of psycholinguistic and syntactic literature. Each of these metrics may explain high or low acceptability through different mechanisms and principles embedded in syntactic parsing. A key criterion in their selection was ensuring that none of the metrics exhibited a correlation higher than 0.5 with any of the others, thereby maintaining their independence in terms of explanatory power.

4.1 Number of Displaced Words

This metric is our operational adaptation of the optimality-theoretic model proposed by Kallestinova (2007). While Kallestinova did not formulate an explicit metric, her analysis draws on the Linearity-IO constraint (McCarthy and Prince, 1995), which penalizes deviations from canonical word order. Within this framework, each displacement from the base order is interpreted as a violation that increases processing cost and reduces acceptability.

To translate this into a measurable form, we defined the metric as the number of displaced words relative to the canonical [Subj Aux V Adj Obj] order. For instance, [Subj Aux V Adj Obj] incurs 0 displacements, while [Subj Aux Obj Adj V] incurs 2, as both the object and the verb are misaligned with their canonical positions. The metric does not consider the direction or length of the displacement – only the number of misordered elements.

4.2 Total Path of Displaced Words

This metric extends Kallestinova's approach by focusing exclusively on the linear distance of displacement. The rationale for this metric lies in the idea that longer displacements require greater cognitive effort for processing and greater distortion from canonicity thereby reducing sentence acceptability.

The metric evaluates the cumulative linear distance of displacements relative to the canonical [Subj Aux V Adj Obj]. For each displaced element, the distance is calculated as the absolute difference between its ordinal position in the canonical order and its actual position in the sentence. For instance, given the canonical order Subject (1), Aux (2), Verb (3), Adjective (4), Object (5), a sentence like "Subject Aux Verb Object Adjective" would involve the following calculation: $|4-5|+|5-4|=2$. Thus, the higher the total score, the lower the acceptability rating.

4.3 Number of Projectivity Violations

This metric is rooted in dependency grammar and functional approaches to language (Liu et al., 2017). Dependency grammar operates under four fundamental rules as outlined by Robinson (1970):

1. One and only one element is independent.
2. All other elements depend directly on some element.
3. No element depends directly on more than one other.
4. If A depends directly on B and some element C intervenes between them in the linear order of the string, then C must depend directly on A, B, or another intervening element.

The fourth rule defines projectivity, stipulating that dependency arcs must not cross each other or the root node.

Here we followed the Dependency grammar principles applying two separate approaches, as in MDD calculations: one assuming the verb as the root node, based on de Marneffe et al. (2014), and the other assuming tense auxiliary as the root node, as in Groß and Osborne (2015). In our dependency calculation process, each crossing of a dependency arc is counted as one violation (Type I / strong violation) (Lu et al., 2016; Testelets, 2001). Similarly, any crossing between an arc and the root node is counted as one violation (Type II / weak violation). The final metric is the sum of both violation types.

To illustrate, in Figure 1, an auxiliary-root structure adheres to projectivity constraints, resulting in zero violations. In Figure 2, the arcs crossing introduces one strong violation, while the crossing between the arc and the root node (Aux) adds one weak violation, totaling 2 violations. Again, a higher number of violations corresponds to a lower acceptability rating.

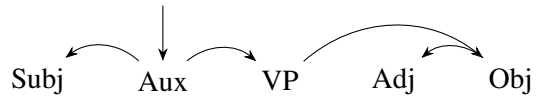

Figure 1: An example of a sentence structure with no projectivity violations.

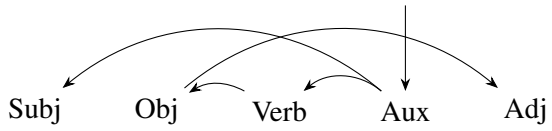

Figure 2: An example of a sentence structure with strong and weak projectivity violations.

4.4 Word Order Penalty Score

Building on the work of (Urošević et al., 1986), the Word Order Penalty Score metric quantifies the cognitive cost of deviations from the canonical SVO structure in Slavic languages. Unlike metrics based on projectivity violations, this metric focuses on the linear disruption of canonically adjacent syntactic pairs, regardless of whether their dependency arcs cross others or the root. Experimental findings show that noncanonical word orders (e.g., VSO, OSV) are generally processed more slowly than canonical ones. This suggests that unlicensed deviations from SVO incur additional processing costs. Urošević and colleagues (1986) attribute this to the rarity of verb- or object-initial sequences in Serbian. Similarly, a corpus-based study by Slioussar and Makarchuk (2022) shows that Russian exhibits a comparable word order frequency distribution, allowing conclusions from Serbo-Croatian to extend to Russian.

The Word Order Penalty Score metric formalizes these observations by assigning penalties to deviations from canonical SVO structure. Sentences retain 0 points if they follow SVO, while noncanonical orders are penalized: object-before-subject (+1), verb-before-subject (+1), object-before-verb (+1). Additional penalties apply for disrupting (an occurence of an intervening element inside the pair) syntactic units such as noun-modifier pairs or auxiliary-verb combinations (+1 each), as these

interruptions increase cognitive cost and memory load by breaking expected grouping patterns (Sekerina, 1999). Thus, the Word Order Penalty Score assigns cumulative penalties for each structural and group-level deviation:

- Object preceding subject: +1 point
- Verb preceding subject: +1 point
- Object preceding verb: +1 point
- Adjective + Object disruption: +1 point
- Auxiliary + Verb disruption: +1 point

4.5 Special Status Residuals and Processing Penalties

This metric integrates elements of formal syntax and psycholinguistics. One frequently observed factor influencing the acceptability of a given word order is its ontological status within the grammar of a particular language, that is, whether it is recognized as a distinct grammatical phenomenon rather than a surface variant. Formal grammatical approaches and empirical studies of word order variability (e.g. Bader and Meng, 1999; Miyamoto and Takahashi, 2002; Sekerina, 1999; Hyönä and Hujanen, 1997) have been especially effective in identifying such configurations. Building on this work, we focus on four phenomena relevant for Slavic languages: (1) Left-Branch Extraction (LBE) in wide sense (extracting Adjective to any leftward position), (2) Noun Scrambling (Object Displacement), (3) violations of canonical clitic or auxiliary placement, captured as NotClitic2 for Serbo-Croatian and NotBudet2 for Russian, and (4) verb topicalization, reflected in the NotSV metric, which detects disruptions of subject-verb adjacency.

A key observation here is that the special discourse status of these constructions may persist even outside of context, allowing them to be perceived as marked permutations and to receive higher acceptability ratings in comparison with their unmarked permuted counterparts. However, when multiple such operations co-occur, processing difficulty increases substantially and tends to override any residual interpretive coherence, resulting in strong acceptability penalties (Novozhilov et al., 2025). This makes it possible to use this metric as a proxy for interactions between processing cost and discourse licensing, capturing how far constructions deviate from both canonical structure and context-sensitive grammatical norms.

The metric assigns penalties to quantify the cumulative processing costs associated with these deviations. Penalties are applied as follows:

- NotBudet2: If present, +1 point; if absent, 0 points.
- Noun Scrambling: If present, +1 point; if absent, 0 points.
- NotSV: If present, +1 point; if absent, 0 points.
- LBE: If present, +1 point; if absent, 0 points.

5 Experiment 1 (Russian)

5.1 Participants

79 adult native Russian speakers (mean age = 30.1) took part in Experiment 1. Participation was voluntary and uncompensated.[1]

5.2 Materials and Procedure

Sentences were constructed by producing all 120 word order permutations of a five-word kernel structure in canonical order: Subj Aux V Adj Obj. The analytical future tense auxiliary *budet* 'be.3SG'/ *budut* 'be.3PL' was used in all sentences. The subject was a noun in the nominative case, and the adjective unambiguously modified the direct object. Subject and object features were systematically varied in animacy, gender, and number. For instance, if the subject was animate, feminine, and singular, the object was inanimate, masculine, and plural, and vice versa. 120 lexical content variants (lexicalizations) were created.

Lexicalizations were distributed across six experimental lists, each containing 20 word orders.[2] Each order was represented by six different lexicalizations, totaling 120 sentences per list. The distribution was randomized.

Participants accessed the experiment via the PCIbex platform (Zehr and Schwarz, 2018). Before starting, they completed a brief demographic questionnaire (age, gender, education level, native language) and signed consent forms. Participants rated the sentences for acceptability on a 5-point

[1]The Russian and Serbo-Croatian experiments were approved by the Ethics Committee of the University of Nova Gorica, protocol no. 4/2024-6.

[2]In our design, participants were presented with one sentence at a time; all preceding and subsequent sentences served as fillers relative to that target. The randomization procedure ensured that different lexicalizations of the same word order never appeared consecutively. Lexical variation was used throughout to maximize filler-like effects. We excluded from analysis all trials in which participants rated the canonical (SVO) sentence below 4, which served as a baseline check for syntactic norm adherence and task engagement. We opted for this design because using actual fillers in the context of studying all possible word order permutations of a five-word long kernel sentence would result in an enormous logistical problem or an unfeasibly large number of trials.

Likert scale in a speeded-acceptability task, with a 7-second limit per sentence.[3] After every 20 sentences, participants could take a short break. Sentences were presented one by one on the computer screen and participants typed numerical acceptability responses. Participation was restricted to PCs and laptops.

6 Experiment 2 (Serbo-Croatian)

6.1 Participants

118 adult self-reported native speakers of Serbo-Croatian participated in the study (mean age=33). Participation was voluntary and uncompensated.

6.2 Materials and Procedure

In Experiment 2 the structure of kernel sentences was the same as in Experiment 1, but only 3 different sentence lexicalizations were used resulting in 3 experimental lists containing the set of all 120 permutations of a single lexicalization each. In all sentences, the grammatical subjects were animate, while the direct object could be animate or inanimate. Gender and number features of the subjects and objects varied across stimuli. All sentences were in past tense and had the clitic *je* (be.3SG) or *su* (be.3PL) as an auxiliary verb. There was no time limit for answers and all stimuli sentences were presented at the same time to the participants. 63 participants from our pool evaluated the first experimental list, 40 participants evaluated the second experimental list, and 15 participants evaluated the third one. Other aspects of the experiment design were identical to Experiment 1.

7 Modeling Results: Processing and Grammar Interaction

First, we examined whether any demographic or lexical factors significantly predicted acceptability ratings. A cumulative link mixed model implemented via the ordinal package (Christensen, 2023) in R (R Core Team, 2021) was used for this purpose. The results indicated that only subject animacy was a significant predictor. Consequently, it was added as a covariate to all subsequent models including the null models, that served as baseline for comparison of models. Subject Animacy was included only in Russian related models, since in Serbo-Croatian stimuli all subjects were animate

Model fit was evaluated using likelihood-ratio tests against the null model, with p-values adjusted via the Holm–Bonferroni correction. Table 1 presents the results of model comparisons for Russian, ranked by AIC (Akaike, 1970) and pseudo-R^2 (McFadden, 1974). The models with the respective tested metrics were numerically coded in the table as follows:

1. Projectivity violations (verb as root),
2. Projectivity violations (auxiliary as root),
3. MDD (auxiliary as root),
4. Word Order Penalty Score,
5. MDD (verb as root),
6. Number of Displaced Words,
7. Total Path Length,
8. Residuals and Processing Penalties.

As shown in Table 1, the best-fitting predictors are projectivity-based metrics (with either verb or auxiliary as root) and MDD in both variants. Although several other models also differ significantly from the null, AIC differences of over 200 strongly favor the top-ranked models, providing clear evidence of superior model fit.

For Serbo-Croatian, we first examined the effect of Wackernagel Law violations (NotClitic2) by comparing the subset of sentences containing such violations to the rest of the dataset. A cumulative link mixed model was fitted using the ordinal package in R (Christensen, 2023) with sentence rating as the dependent variable, NotClitic2 as a fixed effect, and random intercepts for lexicalization and participant (subject_id). Sentences with clitic-placement violations were rated significantly lower ($p<0.001$), with a mean rating of 1.8 compared to 3.44 for the rest of the dataset. All these findings were further confirmed through cross-validation, following (Barth and Kapatsinski, 2018).

The results for Serbo-Croatian highlight several key points. First, the goodness-of-fit patterns across metrics partially align with those observed for Russian. However, the metric based on projectivity violations with the verb as root clearly outperforms all others, with an AIC advantage of over 250 points compared to the next best metric, MDD (verb as root). In contrast, projectivity (auxiliary as root) and MDD (auxiliary as root) fail to reach significance thresholds. This may be attributed to the fact that, after subsetting, we only have sentences with auxiliaries remaining in a fixed position, which results in lower overall word order

[3] We opted for the speeded acceptability task in order to (a) avoid potential satiation effects and (b) explore its advantage as better reflecting initial parsing difficulty (Sprouse, 2008; Weskott and Fanselow, 2011)

Model	AIC	ps.-R^2	p-value	p-bonf
Null	24952.8	NA	NA	NA
1	24420.5	0.02	<0.001	<0.001
2	24536.6	0.01	<0.001	<0.001
3	24538.5	0.01	<0.001	<0.001
4	24739.4	0.009	<0.001	<0.001
5	24739.8	0.009	<0.001	<0.001
6	24934.9	<0.001	<0.01	<0.01
7	24945.9	<0.001	<0.05	0.023
8	24953.6	<0.001	0.3	1.00

Table 1: Comparison of metrics' fit for Russian data

Model	AIC	ps.-R^2	p-value	p-bonf
Null	7478.68	NA	NA	NA
1	7220.93	0.04	<0.001	<0.001
2	7479.43	<0.001	0.26	1.00
3	7480.15	<0.001	0.47	1.00
4	7480.47	<0.001	0.64	1.00
5	7401.95	0.01	<0.001	<0.001
6	7440.82	0.005	<0.01	<0.01
7	7404.36	0.01	<0.001	<0.001
8	7451.93	0.004	<0.01	<0.01

Table 2: Comparison of metrics' fit for Serbo-Croatian data, clitic-second condition.

variability than in Russian. In contrast, main verbs in Serbo-Croatian exhibit greater positional flexibility, making the verb-based projectivity metric more sensitive and explanatory in this context.

These findings emphasise that language-specific grammatical constraints must be modelled explicitly in "free" word-order systems. When discourse cues are deactivated (as in out-of-context judgements), ease of processing alone cannot compensate for categorical violations. Omitting grammar–processing interactions can therefore blur metric performance (see Table "AIC_and_R2_SC_data" in our OSF repository). We argue, therefore, that predictive models should include such constraints as independent factors. A promising next step for future work is a weighted framework that gives grammar maximal weight under decontextualised conditions, for example, assigning grammatical constraints 100% of the weight when no contextual licensing is available, then letting processing metrics assume a larger role when discourse support is present.

8 A Closer Look at the Projectivity

While the difference in projectivity-based metrics for Serbo-Croatian has a straightforward explanation, the analogous divergence in Russian presents a conundrum. A projectivity violation is defined based on the number of crossing dependencies of various types. But why should two implementations of the same general principle, with either the verb or the auxiliary as root, differ so markedly in model performance?

We propose that the discrepancy arises not solely from the quantity of crossings, but from the type of dependency relation whose arc performs the crossing, a subtle but crucial factor [4]. Consider the dependency structure when the Verb is taken as root: Verb → Object, Verb → Subject, Verb → Auxiliary, Object → Adjective. Since arcs emanating from the same node cannot intersect, only one arc in this configuration is capable of producing projectivity violations of both types – Object → Adjective.

Now consider the structure when the Auxiliary is taken as root: Auxiliary → Subject, Auxiliary → Verb, Verb → Object, Object → Adjective. In this configuration, two dependencies: Verb → Object and Object → Adjective – are structurally eligible to cross other arcs. Thus, the auxiliary-root model includes a distinct type of projectivity violation not present in the verb-root model. It is plausible that these dependency types differ in processing cost, leading to divergent model performance despite comparable violation counts.

To evaluate this hypothesis, we conducted three additional statistical analyses for Russian. First, we removed all sentences in which projectivity violations involved only VO-type dependencies (Verb → Object), and compared Aux-root and V-root models once again. The resulting AICs equaled 19035,4 for Aux-as-root, and 19050,5 for V-as-root. We assessed the robustness of this difference using a non-parametric bootstrap. Given that the two models under comparison are non-nested and AIC values can be sensitive to data perturbation, we applied a bootstrapping procedure to evaluate whether the observed $\Delta\mathrm{AIC}$ is stable across resampled datasets. The original dataset consisted of 9123 observations. For each of 1000 bootstrap iterations, we resampled the full dataset with replacement. The mean difference in AICs was 14.2

[4] We are grateful to an anonymous reviewer for raising this possibility.

in favor of the Aux model. However, the 95% confidence interval [-29.43, 45.59] included 0 and p = 0.712, indicating that the observed difference is not statistically robust. We therefore conclude that the apparent advantage of the Aux model is not reliably supported across resampled datasets.

In addition, these results indicate that a VO-dependency is easier to process than an Object-Adjective(AO) dependency. To further test this hypothesis and also test the significance of violation type (weak vs strong) we ran a cumulative link mixed-effects model on acceptability values testing an interaction term dependency_type*violation_type, where dependency type ranges over VO, AO, or both and violation type values were weak, strong and weak+strong (since a violation can be weak and strong at the same time). Random effects included participant and lexicalization. This was followed by pairwise comparisons with Tukey correction, whose results are reported in Table 3.[5]

Comparison	β	SE	z	p
VO.s-VO.w	-0.085	0.12	-0.68	0.999
VO.s-VO.ws	-0.38	0.14	-2.8	0.113
VO.s-AO.s	1.24	0.1	11.7	**<.0001**
VO.s-AO.w	0.61	0.15	3.9	**0.0025**
VO.s-AO.ws	1.02	0.12	8.5	**<.0001**
VO.w-VO.ws	-0.3	0.14	-2.1	0.458
VO.w-AO.s	1.32	0.11	11.7	**<.0001**
VO.w-AO.w	0.69	0.16	4.4	**0.0004**
VO.w-AO.ws	1.11	0.12	9.4	**<.0001**
VO.ws-AO.s	1.62	0.13	12.3	**<.0001**
VO.ws-AO.w	0.99	0.17	5.9	**<.0001**
VO.ws-AO.ws	1.41	0.14	9.9	**<.0001**
AO.s-AO.w	-0.63	0.14	-4.5	**0.0003**
AO.s-AO.ws	-0.21	0.1	-2.1	0.49
AO.w-AO.ws	0.42	0.14	2.9	0.09

Table 3: Pairwise comparisons of dependency and violation types, ($p < 0.05$) are in bold in the text.

Table 3 shows that the factor dependency_type plays a substantial role in shaping acceptability judgments. Specifically, violations caused by VO arcs differ significantly from their AO counterparts: even weak violations, where the AO dependency crosses the root, are rated significantly lower than any VO violation. In addition, the type of violation interacts with the dependency type. In the VO condition, the distinction between weak, strong, and combined violations does not yield significant differences in ratings. In contrast, within AO structures, weak violations are rated significantly higher than strong and marginally than combined violations, although the difference between AO_strong and AO_weak+strong is not statistically significant. Moreover, the effect of weak violation seems to be blurred in mixed types.

Importantly, however, the question of how dependency type and violation type work together to predict acceptability is still far from resolved. Among the 120 permuted sentences, some involved mixed violations – for instance, [Obj Subj V Aux Adj] includes both an AO-arc (a weak violation) and a VO-arc (a strong violation). Investigating how such combinations of violations affect acceptability represents a promising direction for future research, as the effects appear to be cumulative and potentially multidirectional.

Finally, to further disentangle the influence of violation number from dependency type, we conducted an ordinal regression on a subset of sentences containing only strong violations involving the AO dependency. The model used the number of strong violations as the only predictor (with the same random effects structure). The results showed that sentences with a single strong violation were rated significantly higher than those with two violations (p = 0.0007), strengthening the conjecture that the number of violations independently contributes to acceptability judgments.

9 The Interaction of MDD and Projectivity

The final question is whether MDD and projectivity violation metrics capture overlapping or complementary aspects of word order processing—i.e., whether combining them improves predictive accuracy across languages compared to using either metric alone. To address this, we recalculated AIC and pseudo-R^2 values using projectivity violations as the baseline, given that earlier tests indicated that MDD alone performed worse. For consistency, and because the bootstrap analysis showed no significant difference between Verb-as-root and Aux-as-root calculations, we adopted the Verb-as-root configuration for both languages. The resulting model comparisons are shown in Tables 4 and 5 for Russian and Serbo-Croatian, respectively.

[5]We excluded comparisons involving the "both" condition from the presentation for reasons of brevity. The full model output is available on the OSF

Model Name	AIC	Pr(>Chisq)
Proj. (V-Root)	24420.54	NA
Proj. + MDD	24330.29	<0.001

Table 4: Comparison of Mixed Models: MDD vs. Projectivity + MDD (Verb as Root), Russian

Model Name	AIC	Pr(>Chisq)
Proj. (V-Root)	7220.93	NA
Proj. + MDD	7192.32	<0.001

Table 5: Comparison of Mixed Models: MDD vs. Projectivity + MDD (Verb Root), Serbo-Croatian

These findings indicate that MDD and projectivity violation metrics are not redundant, but rather complementary. Their combined use enhances the model's ability to account for the distribution of acceptability ratings in free word order languages. Importantly, the joint model demonstrates greater descriptive and predictive power.

To further illustrate this, consider two example sentence types: *Adj Aux Subj Obj V* and *V Aux Adj Obj Subj*. Both have MDD = 2.25, yet differ in projectivity violations (2 vs. 0) and acceptability ratings (2.33 vs. 2.97). The ordinal regression showed significant differences in their ratings, $p < 0.05$. This suggests that the two metrics capture distinct cognitive pressures: MDD reflects general processing economy, while projectivity violations relate to structural predictability and locality. Together, they offer a more comprehensive account of how sentence structure is evaluated during comprehension.

10 General Discussion

This study yields several key findings. First, the best-performing model was the one that combined Mean Dependency Distance (MDD) and projectivity violation metrics, outperforming models that used either metric alone. This finding broadens the scope for future inquiry across different theoretical frameworks, including those within the MDD/MDDP tradition (Boston et al., 2011; Ferrer-i Cancho, 2004; Gibson, 1998; Futrell et al., 2015; Gibson, 2000) as well as those focusing on projectivity and its violations (Ferrer-i Cancho, 2017; Yadav et al., 2020, 2022).

Second, our findings suggest that MDD and projectivity encode distinct cognitive mechanisms involved in syntactic parsing. MDD indexes the memory load of cue-based retrieval: longer linear distances force the parser to keep more items active, increasing interference and cost. Projectivity violations, by contrast, disrupt a stack-based incremental parser, penalizing non-projective structures where immediate attachment is not possible (Frazier and Fodor, 1978; Frazier, 1979). In transition-based parsing, projective dependency trees are precisely those that can be derived with a single push-down stack, whereas non-projective trees need extra SWAP operations or an auxiliary stack (Nivre, 2003, 2009). Our final regression model further supports this interpretation. Further research is needed to compare these mechanisms directly, particularly in cases where the two principles overlap or diverge in their predictions.

Third, our data indicate that the type of violation matters, but only for certain dependency types. The contrast we observed between AO and VO dependencies may point to effects of structural embeddedness that may increase the cognitive cost of disrupting the canonical configuration. In both dependency and phrase-structure based grammars, the AO pair is structurally more embedded than the VO relation. Again, targeted studies are required to test this hypothesis explicitly.

Finally, our modeling results underscore the complex interaction of multiple factors contributing to acceptability. Grammatical constraints appear to carry the greatest weight, followed by projectivity violations. Within the latter, it would be valuable to further explore how dependency type, violation type, and number of violations contribute independently and interactively. The third layer is MDD, which acts as a general processing principle affecting sentence structure evaluation. Understanding how these layers interact and compete within the acceptability space is a promising direction for future formal and experimental modeling.

Limitations

One key limitation of this study is that it relies on acceptability judgments, which, while informative, provide only an indirect measure of real-time processing. Also, although two processing principles were modeled (MDD and projectivity violations), other cognitive and discourse-level factors, such as information structure, prosody, or thematic prominence, were not explicitly controlled or integrated into the models, leaving open questions about their interaction with structural constraints.

Acknowledgments and Data Availability

We thank three anonymous reviewers of this paper for their useful feedback. This research has received funding from the Slovenian Research and Innovation Agency (ARIS) under project no. J6-4615. All materials and experimental results from this study are available in our OSF repository at https://osf.io/7p29c/?view_only=ce1cc0a390a24865b76558f9974dceef.

References

Hirotugu Akaike. 1970. Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, 22:203–217.

Markus Bader and Michael Meng. 1999. Subject-object Ambiguities in German Embedded Clauses: An Across-the-Board Comparison. *Journal of Psycholinguistic Research*, 28(2):121–143.

John F. Bailyn. 1995. *A configurational approach to Russian "free" word order*. Ph.D. thesis, Cornell University, Ithaca.

Danielle Barth and Vsevolod Kapatsinski. 2018. Evaluating Logistic Mixed-Effects Models of Corpus-Linguistic Data in Light of Lexical Diffusion. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed-Effects Regression Models in Linguistics. Quantitative Methods in the Humanities and Social Sciences*, 1 edition, volume 1, pages 99–116. Springer International Publishing, Cham, Switzerland.

Otto Behaghel. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25(1909):110–142.

Marisa F. Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Željko Bošković. 2005. Left branch extraction, structure of np, and scrambling. In Joachim Sabel and Mamoru Saito, editors, *The free word order phenomenon: Its syntactic sources and diversity*, pages 13–73. Mouton de Gruyter, Berlin. Accessed 13 June 2025.

Željko Bošković. 2001. *On the Nature of the Syntax-Phonology Interface: Cliticization and Related Phenomena*, volume 60 of *North Holland Linguistic Series: Linguistic Variations*. Brill, Leiden. Accessed 13 June 2025.

Hyo-Woon Choi. 2007. Length and order: A corpus study of Korean dative-accusative construction. *Discourse and Cognition*, 14(3):207–227. Accessed 13 June 2025.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press.

Rune H. B. Christensen. 2023. *ordinal–Regression Models for Ordinal Data*. R package version 2023.12-4.1.

Meredyth Daneman and Patricia A. Carpenter. 1980. Individual Differences in Working Memory and Reading. *Journal of Verbal Learning & Verbal Behavior*, 19(4):450–466.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silvaire, Katrin Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A Cross-Linguistic Typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Accessed 13 June 2025.

Ramon Ferrer-i Cancho. 2004. Euclidean distance between syntactically linked words. *Phys. Rev. E*, 70(5):056135.

Ramon Ferrer-i Cancho. 2017. The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach. *Preprint*, arXiv:1705.09932. ArXiv preprint, version 3.

Lyn Frazier. 1979. *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. thesis, University of Connecticut.

Lyn Frazier and Janet Dean Fodor. 1978. The Sausage Machine: A New Two-Stage Parsing Model. *Cognition*, 6(4):291–325.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336–10341.

Edward Gibson. 1998. Linguistic Complexity: Locality of Syntactic Dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261–290.

Thomas Groß and Timothy Osborne. 2015. The Dependency Status of Function Words: Auxiliaries. In *International Conference on Dependency Linguistics*.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*, volume 73 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.

Richard Hudson. 1995. Measuring syntactic difficulty. Unpublished paper. Available at http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf. Accessed 15 June 2025.

Jukka Hyönä and Heli Hujanen. 1997. Effects of Case Marking and Word Order on Sentence Parsing in Finnish: An Eye Fixation Analysis. *The Quarterly Journal of Experimental Psychology Section A*, 50(4):841–858.

Elena D. Kallestinova. 2007. *Aspects of Word Order in Russian*. Ph.D. thesis, University of Iowa, Iowa City. Accessed 8 December 2024.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Qun Lu, Chunshan Xu, and Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages? *Complexity*, 21:33–41.

John J. McCarthy and Alan Prince. 1995. Faithfulness and reduplicative identity. In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. University of Massachusetts.

Daniel McFadden. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In Paul Zarembka, editor, *Economic Theory and Mathematical Economics*, pages 105–142. Academic Press, New York. Accessed 13 June 2025.

Igor' A. Mel'čuk. 2009. Dependency in natural language. In Alain Polguère and Igor' A. Mel'čuk, editors, *Dependency in Linguistic Description*, pages 1–110. John Benjamins Publishing.

Edson T. Miyamoto and Shoichi Takahashi. 2002. The Processing of Wh-Phrases and Interrogative Complementizers in Japanese. In *Japanese/Korean Linguistics*, volume 10, pages 62–75.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of IWPT-2003*, pages 149–160, Nancy, France.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of ACL-IJCNLP 2009*, pages 351–359, Singapore.

Artem Novozhilov, Kirill Chuprinko, and Arthur Stepanov. 2025. Dense sentence sets induce an anchor–and–baseline strategy in likert scale acceptability judgments. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*. To appear.

Charles A. Perfetti and Alan M. Lesgold. 1977. Discourse Comprehension and Sources of Individual Differences. In Marcel A. Just and Patricia Daneman, editors, *Discourse Comprehension and Sources of Individual Differences*, pages 141–183. Pittsburgh University.

R Core Team. 2021. R: A Language and Environment for Statistical Computing.

Jane J. Robinson. 1970. Dependency Structures and Transformational Rules. *Language*, 46(2):259–285.

Irene Ros, Marta Santesteban, Kazuko Fukumura, and Itziar Laka. 2015. Aiming at shorter dependencies: The role of agreement morphology. *Language, Cognition and Neuroscience*, 30(9):1156–1174.

Irina A. Sekerina. 1999. The Scrambling Complexity Hypothesis and Processing of Split Scrambling Constructions in Russian. *Journal of Slavic Linguistics*, 7(2):265–304.

Natalia Slioussar and Ilya Makarchuk. 2022. SOV in Russian: A Corpus Study. *Journal of Slavic Linguistics*, 30(3):1–14.

Jon Sprouse. 2008. The effect of task demands on acceptability judgments. *Journal of Linguistics*, 44(2):387–408.

Adrian Staub, Charles Clifton, and Lyn Frazier. 2006. Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language*, 54(3):389–406.

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

Jakov G. Testelets. 2001. *Introduction to General Syntax [Vvedenie v obshchiy sintaksis]*. Russian State University for the Humanities.

Zoran Urošević, Claudia Carello, Milan D. Savić, Georgije Lukatela, and Michael T. Turvey. 1986. Some word-order effects in Serbo-Croat. *Language and Speech*, 29(2):177–195.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Weskott and Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language*, 87(2):249–273.

Himanshu Yadav, Samar Husain, and Richard Futrell. 2022. Assessing Corpus Evidence for Formal and Psycholinguistic Constraints on Nonprojectivity. *Computational Linguistics*, 48(2):375–401.

Himanshu Yadav, Ashwini Vaidya, Vishakha Shukla, and Samar Husain. 2020. Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science*, 44(4):e12822.

Jeremy Zehr and Florian Schwarz. 2018. PennController for Internet-Based Experiments (IBEX).

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual Evaluation of Linguistic Acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.