**It's not What You Say but How You Say It: Evidence from Russian Shows Robust Effects of the Structural Prior on Noisy Channel Inferences**

Corresponding Author: Moshe Poliak, MIT, moshepol@mit.edu

**Abstract**

Under the noisy-channel framework of language comprehension, comprehenders infer the speaker's intended meaning by integrating the perceived utterance with their knowledge of the language, the world, and the kinds of errors that can occur in communication. Previous research has shown that, when sentences are improbable under the meaning prior (implausible sentences), participants often interpret them non-literally. The rate of non-literal interpretation is higher when the errors that could have transformed the intended utterance into the perceived utterance are more likely. However, previous experiments on noisy channel processing mostly relied on implausible sentences, and it is unclear whether participants' non-literal interpretations were evidence of noisy channel processing or the result of trying to conform to the experimenter's expectations in an experiment with nonsensical sentences. In the current study, we used the unique properties of Russian, an understudied language in the psycholinguistics literature, to test noisy-channel comprehension using only simple plausible sentences. The prior plausibility of sentences was tied only to their word order; SVO sentences were more probable under the structural prior than OVS sentences. In two experiments, we show that participants often interpret OVS sentences non-literally, and the probability of non-literal interpretations depended on the Levenshtein distance between the perceived sentence and the (potentially intended) SVO version of the sentence. The results show that the structural prior guides people's final interpretation, independent of the presence of semantic implausibility.

**Keywords: Sentence Processing; Noisy Channel Processing; Structural Prior; Russian**

## Introduction

We live and act in a world rife with ambiguity, constantly needing to make decisions under uncertainty (Tversky & Kahneman, 1974). Such decisions may be matters of life-and-death: for example, to successfully cross the street, we need to rapidly compute the probability of being hit by a car, even though we can only *estimate* how far away it is or how fast it is moving based on our visual and auditory input. An area where we constantly reason under uncertainty, although usually without deadly consequences, is language. Past work has shown how probabilistic inference may be used in many aspects of language use including word learning, speech perception, and pragmatic interpretation (e.g., Clayards et al., 2008; Goodman & Frank, 2016; Xu & Tenenbaum, 2007). Here, we deal with uncertainty in sentence processing. In everyday language use, the language that listeners hear or see is full of noise. Noise manifests in phenomena like memory failures, mishearings, speech errors, typographical errors, ad-hoc sentence repairs (e.g., "We should to the bea… to the pool, we should go to the pool."), and environmental noise (e.g., when talking in a club with loud music). Despite this ever-present noise, communication often unfolds successfully and effortlessly.

How do humans manage to communicate so effectively despite the noise in language? According to the *noisy channel* processing model, to overcome ambiguity and noise in language, comprehenders integrate the perceived input (sounds, signs, or written forms) with their expectations about the meanings that the speaker may convey (Gibson et al., 2013; Levy, 2008; Levy et al., 2009). We build upon this research and ask whether, in pursuit of understanding the message, we use not only our expectations about *what* the speaker will say, but also our expectations about *how* the speaker will say it.

**The Noisy Channel Model**

Participants who read grammatically well-formed but implausible sentences sometimes interpret them non-literally[1] as a more plausible alternative, especially when the implausible sentence is similar to a plausible alternative (Ferreira et al., 2002; Gibson et al., 2013; Traxler, 2014). For example, "The mother gave the candle the daughter," which could have resulted from the more plausible alternative "The mother gave the candle *to* the daughter," is often interpreted as the more plausible choice, as indicated by participants answering "yes" to the question "Did the daughter receive something?"

Following Shannon (1949), recent psycholinguistic models have argued that the process of inferring the speaker's intended meaning from noisy input can be formalized as Bayesian reasoning (Gibson et al., 2013; Levy, 2008; Levy et al., 2009). According to this noisy-channel framework (**Eq. 1**), the probability of the intended sentence, $S_i$, given the perceived sentence, $S_p$, is proportional to the prior probability of the intended sentence given knowledge of the world and the language, $P(S_i)$, and the probability that $S_i$ would be corrupted to $S_p$ during transmission, $P(S_p|S_i)$. In other words, $P(S_i)$ represents how likely an utterance is, while $P(S_p|S_i)$ represents how likely various errors are under the comprehender's noise model.

$$P(S_i|S_p) \propto P(S_p|S_i) * P(S_i) \quad (1)$$

The noisy channel framework uses the prior and likelihood functions to explain how, sometimes, interpreting a sentence non-literally may help overcome noise and help communication. In many—perhaps most—instances, we will interpret an utterance literally as

[1] We intentionally refer to interpretations as "literal" and "non-literal" rather than "correct" and "incorrect" because, in our view, an interpretation is "correct" if it successfully recovers the meaning that the producer intended, which might apply to either literal or non-literal interpretations. On the other hand, an interpretation is literal if it reflects the conventional compositional meaning of the sentence (i.e., using the language's lexicon and grammar to interpret the sentence). This is the relevant dimension for the comprehension questions in our study.

presented. If the speaker intends and produces $S_i$ = "The mother gave the candle to the daughter," then we are most likely to perceive utterance $S_p$= "The mother gave the candle to the daughter", more than any other possible alternative. A non-literal interpretation is expected only when the literal interpretation of the perceived utterance is odd in some way, for example, if one perceives a sentence like "The mother gave the candle the daughter." The literal interpretation here is unlikely because it states that the mother is giving a person (the daughter) to an inanimate object (a candle), an unlikely event with a low prior probability, $P(S_i)$. The noisy channel framework proposes that such sentences may be resolved during processing by assuming that the perceived sentence was somehow corrupted, and, in fact, a different sentence was intended. How likely it is that utterance $S_p$ was perceived if $S_i$ was intended is quantified using the likelihood term, $P(S_p|S_i)$. Consequently, the aim of processing is to find utterance $S_i$ that maximized the product of the prior, $P(S_i)$, and the likelihood, $P(S_p|S_i)$.

Gibson et al. (2013) proposed that the likelihood of an utterance, $P(S_p|S_i)$, is proportional to the Levenshtein distance (Levenshtein, 1966) between the perceived and the intended utterances: the more edits separate the intended and produced utterances, the lower the likelihood. For example, the plausible sentence "The girl kicked the ball" requires two word-level edits to produce the implausible sentence "The girl was kicked by the ball." As such, the likelihood that the first sentence was intended when the second was perceived is low. This contrasts with materials like "The mother gave the candle (to) the daughter", which only requires assuming one edit (a deletion of "to") to change the plausible, potentially intended, sentence into the implausible, perceived sentence. Indeed, Gibson et al. (2013) show that participants almost always interpret sentences like "the girl was kicked by the ball" literally, even though they are implausible; and they show that sentences like "The mother gave the candle the daughter" are

often interpreted as if the word “to” were deleted, inferring a more plausible meaning in this case.

Furthermore, Gibson et al. (2013) proposed that different *types* of edits may have different likelihoods. In particular, edits may be categorized into deletions and insertions. For deletions, a plausible sentence could result in an implausible sentence if a part of it is dropped (e.g, “The mother gave the candle (to) the daughter”). For insertions, a plausible sentence could result in an implausible sentence if an element is added to it erroneously (e.g., “The mother gave the girl to the candle”). Gibson et al. (2013) proposed that deletions are more likely production errors than insertions. For a deletion, a single element from the sentence is selected; but for an insertion, a single element from the entire vocabulary undergoes the edit, resulting in a reduced probability that any specific word was inserted. Other types of edits, like when one word is substituted for another word, can be seen as a combination of an insertion and a deletion, and, accordingly, have the cost of both a deletion and an insertion, under the hypothesis of Gibson et al. (2013).

The overall likelihood of errors can vary as well, depending on the reliability of the incoming signal. The less reliable the signal, the higher the likelihood of errors and the more comprehenders rely on their priors rather than the observed signal. In Gibson et al. (2013), filler materials were manipulated such that, for some participants, half the filler materials were ungrammatical (e.g., “A legislator lied to the consultant a new bill.”). Participants who observed ungrammatical filler sentences interpreted implausible critical sentences non-literally more often than participants who observed the grammatical filler sentences only (Gibson et al., 2013). This suggests that noisier environments cause participants to rely more on their prior than they would in a less noisy environment.

Comprehenders consider not only word-level edits to the perceived sentence, but also character-level (or segment-level) edits (Keshev & Meltzer-Asscher, 2021; Levy et al., 2009; Ryskin et al., 2021). Ryskin et al. (2021) showed that, in the presence of a contextually low-probability word, the probability of inferring that a more plausible alternative word was intended (as opposed to interpreting the word literally and inferring no corruption) can be indexed by the N400 and P600 ERP components. The N400 is often seen as an index of change in the comprehender's semantic representation (Kutas & Federmeier, 2011; Rabovsky et al, 2018). In contrast, the P600 is less well understood, but it has been tied to violations of form (Osterhout & Holcomb, 1992; Münte et al., 1998) or nonsensical input that requires repair (Kuperberg et al., 2020), and is more generally related to the family of P300 components that are involved in detection of low probability events (Leckey & Federmeier (2019). In line with these accounts, Ryskin et al. (2021) showed that, when readers see a word that is implausible in context and has an available *alternative* (e.g., "The storyteller could turn any incident into an amusing antidote/*anecdote*"), the N400 is reduced and the P600 is increased relative to a case where no noisy-channel inference is likely. This reflects that, when a noisy channel inference occurs, the interpretation of the sentence is in terms of the meaning of the more plausible alternative. Further, the amplitude of the P600 was negatively correlated with the Levenshtein distance between the perceived implausible word and the plausible alternative word, suggesting that it indexes the probability of noisy-channel inference taking place. This indicates that participants may consider similar alternative utterances based on noise corruptions across multiple levels of granularity (e.g., characters, segments, words, and multi-word phrases).

Varying the prior probability of an utterance also affects the rate of non-literal interpretation. The prior of an utterance, $P(S_i)$, represents how likely an intended sentence is

given the knowledge of the comprehender. Priors may differ across contexts and be influenced by many sources of information. In a similar paradigm to Gibson et al. (2013), Nathaniel et al. (2018) showed participants plausible and implausible sentences preceded by context sentences. In the experimental condition, the context sentences were related to the target sentence, while in the control condition they were unrelated. For example, in the experimental condition, the implausible sentence "the girl threw the boy to the apple" could appear in the context "The boy and the girl went apple picking together. The girl picked an apple that the boy wanted." They found that participants were more likely to interpret an implausible sentence non-literally if the non-literal (plausible) interpretation was supported by the context, as in the example above. Having additional semantic information (e.g., that the boy wanted an apple that the girl had), increased the prior probability of the intended sentence (i.e., it is very likely that the girl would throw the apple to the boy). Moreover, the prior is tuned to the context. The more implausible sentences are encountered, the more probable implausible sentences become. Specifically, in Gibson et al. (2013), increasing the proportion of implausible sentences increased participants' proportion of literal interpretations. We can interpret this to mean that, when participants notice many implausible sentences, the likelihood of implausible sentences to be intentional, not erroneous, increases.

**The Structural Prior**

While previous work showed that the rate of non-literal inference can be manipulated by varying the semantic plausibility of a sentence, we ask whether the rate of inferences can be manipulated by varying the *form* likelihood of the sentence—the structural prior—while preserving a plausible meaning. Initial evidence that the structural prior may affect the inference

rate can be found in Gibson et al. (2013), with respect to the implausible locative inversion materials that they investigated: e.g., “The table jumped onto the cat” or “Onto the cat jumped a table.” Each of these can be made plausible by changing the position of the word “onto”. Whereas this was the case for materials using a common word order “The table jumped onto the cat” (94.1% literal; 5.9% inference), there was a surprisingly high proportion of non-literal inference in the low-frequency structure “Onto the cat jumped the table” (84.8% literal; 15.2% inference). This discrepancy can be explained in terms of the structural prior: the frequency of a construction in the language. The sentence “Onto the cat jumped the table” has an infrequent locative inversion word order, and it consequently may be less likely to be intended than “The table jumped onto the cat”, despite the same number and type of edits separating them from their plausible alternatives.

Poppels and Levy (2016) extended the design in Gibson et al. (2013) by manipulating not only the plausibility of sentences, but also the canonicality of sentences. They define a construction to be canonical if it is frequent in the language. Poppels and Levy (2016) first showed that certain sequences of post-verbal prepositional phrases were more common than others. E.g., a “from” phrase usually precedes a “to” phrase following a verb like “fell”, as in “The package fell from the table to the floor.” They then showed that participants were more likely to interpret a sentence non-literally if it had a non-canonical construction, such as “The package fell **to the table** from the floor,” than a sentence that had a canonical construction, such as “The package fell from the floor **to the table**.”

In another study, using both English and Mandarin, noisy channel inferences were shown to increase when word order was non-canonical (Liu et al., 2020). Participants were presented with plausible and implausible sentences, in either the canonical SVO word order, or the non-

canonical OSV word order (e.g., “The trash, the boy threw”). Liu et al. (2020) showed that sentences with OSV word order were more likely to be interpreted non-literally than sentences with SVO word order, especially in implausible sentences.

Further exploring the role of the structural prior, Keshev and Meltzer-Asscher (2021) used several tasks with online and offline processing in Hebrew. In their offline sentence completion task, Keshev & Meltzer-Asscher (2021) show that participants may opt to complete a sentence with an agreement error in order to avoid non-canonical word order. In the task, participants were presented with a preamble that included a sentence with a beginning of a relative clause (e.g., “we liked the pupil that despite the concerns found…”). They manipulated whether the modified noun (pupil) was singular or plural, and whether the verb (found) agreed with it or not. When the modified noun was plural and the verb did not agree with it, the grammatically correct completion would require a post-verbal subject, but this structure is rare in Hebrew. Rather than producing this grammatical but rare structure, participants often completed the clause as if the verb agreed with the modified noun, resulting in a more common structure but a grammatically incorrect sentence. This finding suggests that when the prior probability of a sentence structure is very low, participants are less likely to produce it faithfully, likely because they posit that an earlier portion of the sentence was corrupted by noise.

The structural prior has also been shown to exert its influence in real time during language processing (Keshev & Meltzer-Asscher, 2021; Levy et al., 2009, but see Cutter et al., 2022). Building on a study by Tabor et al. (2004), Levy et al. (2009) showed participants sentences like “The coach smiled at the player tossed the frisbee,” in an eye-tracking experiment. When participants’ gaze reached the verb “tossed,” they slowed down and often made regressive eye movements to the preposition “at.” However, when presented with

similar sentences with words like “toward” instead of “at,“ regressive eye movements were significantly reduced. Levy et al. (2009) propose an explanation: a reduced relative clause with the verb “tossed” is less likely under the structural prior than a simple sentence with the same verb. Therefore, the reader considers the option that the unlikely structure that they received was a result of a more probable alternative structure that may have been intended but was corrupted by noise. Specifically, the word “at” has orthographic and phonological neighbors (e.g., “as”, “and”) that would produce a frequent construction (e.g., two simple coordinated sentences: “the coach smiled and the player tossed the frisbee”). In contrast, the word “toward” has no neighbors, so the likelihood of any potential noise corruption is very low; thus, the probability that an alternative high-frequency sentence structure gave rise to the received input is very low, too. This suggests that even when reading a grammatical and semantically plausible sentence, when the structure had low probability, participants entertained alternative readings of the sentence that required positing a likely corruption (e.g., as → at).

**The Current Study**

The goal of the current study is to investigate the role of the structural prior in noisy channel inferences. We investigate this question in an offline comprehension paradigm, thus probing participants’ final interpretations of the sentence meaning, rather than their production behavior or online processing (cf. Keshev & Meltzer-Asscher, 2021; Ryskin et al., 2021). This approach gets at the core of our question of what information ends up being transmitted as a result of the communicative act. While Poppels and Levy (2016) showed that non-canonical sentences are more likely to be interpreted non-literally, their obtained effect size was small. Liu et al. (2020) found similar results using the rare topicalization construction (e.g., “the trash, the

boy threw"). Both experiments manipulated the structure of the stimuli as well as their plausibility. It is possible that findings from implausible stimuli are limited in generalizability. That is, implausible stimuli may engage cognitive processes somehow differently than language processing in everyday communication. Because the current study uses entirely plausible stimuli, it allows us to evaluate this possibility and investigate behavior in the absence of implausible materials.

In addition, we investigate whether the kinds of edits that Gibson et al. (2013) proposed to be involved in non-literal interpretation of implausible materials are also at play when the structural prior is varied. Following Gibson et al. (2013), the simplest theory is that deletions are most likely, insertions are less likely, and combinations of edits are the least likely in these cases.

To do this, we extend the paradigm in Gibson et al. (2013) to Russian, manipulating the structure of the stimuli while holding plausibility constant. Russian has rich morphology and flexible word order compared to English, allowing us to systematically investigate noisy channel inferences while manipulating the structure of the sentence. While, in English, a sentence like "Selena hugged William" can only be grammatically interpreted such that "Selena" is the subject of the verb "hugged," this is not the case in Russian. Based on the verb conjugation and the case of the first and last noun phrases, "Selena" may be the subject, as in 1.a., or the object, as in 1.b. In 1.a, two factors unambiguously indicate that Selena is the subject of the sentence and William is the object. First, Selena and William are morphologically marked for nominative and accusative cases, respectively. Second, the verb agrees in gender and number with the first noun phrase, Selena, marking it as the subject. Sentence 1.a thus has SVO word order. In example 1.b, the same morphological and agreement factors point to William being the subject, and the word order of the sentence is OVS.

1.

a. Селена обняла Вильяма.
   S$^{j}$el$^{j}$ena-Ø obn$^{j}$al-a Vil$^{j}$jam-a.
   Selena-NOM hugged-FEM William-ACC.
   Selena hugged William.

b. Селену обнял Вильям.
   S$^{j}$el$^{j}$en-u obn$^{j}$al-Ø Vil$^{j}$jam-Ø.
   Selena-ACC hugged-MASC William-NOM.
   William hugged Selena.

While word order is flexible in Russian, not all word orders are equally common and the degree of flexibility of the word order depends on morphological marking in the sentence. A corpus study of Russian involving 8,575 clauses revealed that, when the subject and object are morphologically unambiguous (through case markings or verb agreement), the most common word order is SVO (84.18%), followed by OVS (8.99%; Berdicevskis & Piperski, 2020). Example 1a has the most frequent word order, SVO. Although Example 1b has OVS order, it is natural sounding and unambiguous due to case markings and verb agreement. In contrast, when the subject and object are morphologically ambiguous (due to ambiguous case marking and verb agreement), the most common word order, SVO, becomes even more frequent (87.15%), while other word orders, like OVS, become less frequent (7.58% for OVS)[2]. In a majority of cases, even when morphosyntactic information is absent, context and world knowledge suffice for assigning correct agent and and patient roles (Mahowald et al., 2022). For this reason, we generated stimuli that are equally plausible as SVO or OVS sentences, pitting grammaticality against the structural prior while holding the meaning prior constant.

[2] The phenomenon where free word order in a language loses its flexibility in sentences that lack morphological marking is termed *word order freezing* (Bouma, 2011; Jakobson, 1971).

Our study used critical sentences of the form NP V NP, with transitive verbs and noun phrases that consisted of names that do not inflect for case (foreign names such as "Joe" or "Elizabeth"). This resulted in sentences that have only two cues to which noun phrase is the subject and which one is the object: word order and subject-verb agreement. When the verb agreed with the first NP, the resulting sentence was a canonical SVO sentence, since both word order and subject-verb agreement suggested that the first NP is the subject. However, when the verb agreed with the last NP, the two cues were at odds with each other: subject-verb agreement indicated that the last NP is the subject while word order canonicality suggested that the first NP is the subject. In a 3x2 within-participant design, we manipulated the type of edit (deletion, insertion, substitution), and whether the verb agreed with the first or the last NP (see **Table 1**). Furthermore, following Gibson et al. (2013), in Experiment 2 we manipulated between participants whether filler sentences were all grammatical or not.

For every sentence, participants were probed for whether they interpreted the first or the last NP as the subject of the sentence. We call a literal interpretation one that assigns the subject position to the NP that the verb agrees with. In non-canonical sentences, where the verb agrees with the final NP (OVS), participants could interpret the sentence as grammatical and non-canonical (OVS), or canonical and ungrammatical (SVO). When sentences were interpreted canonically and ungrammatically, certain edits to the gender-number suffix of the verb had to be assumed. We classified these edits as deletions, insertions, and substitutions. In a deletion, the agreement morpheme is missing from the perceived sentence, while in an insertion, an agreement morpheme is erroneously present in the perceived sentence. In a substitution, the perceived sentence has an agreement morpheme, but a different morpheme was intended (**Table 1**). In contrast to most previous studies of noisy-channel inference, Russian affords minimal pair

comparisons of edit types: across all conditions, only one character is ever inserted, deleted, or substituted.

Based on the noisy-channel framework research, we make three predictions. First, non-canonical sentences will be interpreted non-literally more often than canonical sentences. This is because, for canonical sentences, the structural prior agrees with the grammatical interpretation of the sentence, while, for non-canonical sentences, the structural prior is at odds with the grammatical interpretation of the sentence. Second, sentences where comprehenders assume deletions will have the highest proportion of non-literal interpretations, followed by insertions, followed by substitutions. Substitution edits are predicted to be the least likely of the three because they assume both that a chunk of the intended utterance was deleted and that an unintended chunk was inserted in the perceived utterance; a substitution can be seen as a composition of a deletion and an insertion. Note that the edits in question are not single-word edits (cf. Gibson et al., 2013) but changes in the suffix of the verb. This ordering is justified by Levenshtein distance and previous work with different types of edits, including edits to bound morphemes (Keshev & Meltzer-Asscher, 2021; Zhan et al., 2017). Third, like Gibson et al. (2013), we predict that ungrammatical fillers will result in a higher rate of non-literal interpretations in general. Following the likelihood assumption in Bayesian reasoning, comprehenders rely more on their world knowledge in noisy environments, regardless of whether the experimental manipulation is to plausibility or the structure of the sentence.

## Methods

We conducted two experiments. Experiment 2 was a preregistered replication and extension of Experiment 1, so the methods are described together.

**Transparency and Openness**

The pre-registration for Experiment 2, as well as the materials, analyses, fitted models, and anonymized data for both experiments are available on the study's OSF page (https://osf.io/8tygf/).

**Ethics Approval**

This work has been ethically approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES), Protocol 403000040, Title: Principles of language processing.

**Materials**

In both experiments, participants were administered a textual questionnaire in Russian with 96 trials (32 critical). Each trial consisted of a sentence-question pair. Critical sentences were identical in both experiments, taking the form of NP V NP, where the NPs consisted of a female name, a male name, or a pair of coordinated male and female names (see **Table 1**). Names were chosen to be unambiguously masculine or feminine (see **Appendix A** for a follow up study showing that readers' perceptions of the gender associations of the names were consistent with our categorization). Each sentence was followed by a question. For example, a sentence like "Чарли увидела Рейчел" ("Charlie saw-fem Rachel") would appear with a question like "Увидел ли Чарли кого-то?" ("Did Charlie see anyone?"). In this sentence, "Rachel" is the subject of the sentence because it is the only NP that can agree with the verb in gender, in spite of "Charlie" being in the canonical subject position (preverbal). Therefore, responding "yes" to the question above would indicate a non-literal interpretation, since it would

reflect an interpretation where Charlie, not Rachel, is the subject of the sentence. For each sentence-question pair, we randomized whether a "yes" answer reflects a literal or a non-literal interpretation (e.g., asking either "Did Charlie see someone?" [yes = non-literal] or "Did Rachel see something?" [yes = literal])[3].

To generate the stimuli, eight female and eight male foreign names that do not inflect for case in Russian (e.g., "Charlie") were selected. Each of the 32 critical items had a unique verb. The verbs were transitive, perfective, and in past tense, resulting in a single unique suffix for male, female, and plural grammatical agreement. This property of the chosen verbs allows for changing a verb's grammatical agreement by deleting, inserting, or substituting a single letter. Names and verbs were counterbalanced across eight lists of critical items (see preregistration for full materials). In each list, each verb would appear only once, each time in a different canonicality-gender order combination (with eight such combinations). The canonical (SVO) and non-canonical (OVS) version of each item differed by either a deletion, an insertion, or a substitution edit. For each item in each list, we randomized whether a "yes" or "no" response indicated a literal interpretation of the sentence. To summarize, there were six types of critical sentences that varied in the associated edit type and whether the verb agreed with the first or last NP. This resulted in two within-participant factors: edit type (deletion, insertion, and substitution; the substitution condition had twice as many sentences because of its two associated gender orders, fp and pf) and canonicality (canonical, non-canonical). In total, each participant observed 32 critical trials, 16 of which were canonical. Within each group of 16 canonical or

[3] Because we used true randomization to determine whether "yes" or "no" indicated literal interpretation, the process resulted in unbalanced numbers of trials across conditions where a "yes" answer indicated literal interpretation. To make sure that our inference is not contingent on whether "yes" or "no" indicated a literal response, we have refitted all the models with this variable as a covariate and saw no changes in inference. The entire outputs are reported in **Appendix B** and are available on OSF.

non-canonical trials, there were 4 deletion (type fm) trials, 4 insertion (type mf) trials, and 8 substitution trials (types pf and fp).

**Table 1. The three edits to get from the canonical construction to the non-canonical construction. The unmarked form of the verb is masculine. The suffixes "-a" or "-i" are added to the verb when the subject NP is feminine or plural, respectively.**

| Edit | Type | Canonical version | Non-canonical version |
|---|---|---|---|
| deletion | fm | Рейчел увидела Чарли.<br>Rachel uvidel-**a** Charlie.<br>Rachel saw-**FEM** Charlie. | Рейчел увидел Чарли.<br>Rachel uvidel-**∅** Charlie.<br>Rachel saw-**MASC** Charlie. |
| insertion | mf | Чарли увидел Рейчел.<br>Charlie uvidel-**∅** Rachel.<br>Charlie saw-**MASC** Rachel. | Чарли увидела Рейчел.<br>Charlie uvidel-**a** Rachel.<br>Charlie saw-**FEM** Rachel. |
| substitution | pf | Чарли и Кейт увидели Рейчел.<br>Charlie i Kate uvidel-**i** Rachel.<br>Charlie and Kate saw-**PL** Rachel. | Чарли и Кейт увидела Рейчел.<br>Charlie i Kate uvidel-**a** Rachel.<br>Charlie and Kate saw-**FEM** Rachel. |
| | fp | Рейчел увидела Чарли и Кейт.<br>Rachel uvidel-**a** Charlie i Kate.<br>Rachel saw-**FEM** Charlie and Kate. | Рейчел увидели Чарли и Кейт.<br>Rachel uvidel-**i** Charlie i Kate.<br>Rachel saw-**PL** Charlie and Kate. |

All filler sentences were grammatical sentences in Experiment 1. Half of the filler sentences were of the form NP V NP, but with names that can be inflected for case, thus allowing more flexibility in word order. The other half were slightly longer simple sentences with two human referents in addition to other objects or modifiers (e.g., "Roman forgot Kirill's

promise."). Comprehension questions for fillers were generated individually for each item, involving the subject and main verb (e.g., "Did Roman forget something?"). Half the filler sentences used Russian names and half used foreign names (all of which inflect for case). For one half of the filler items, the literal response was "yes" and for the other half it was "no." In Experiment 2 only, we varied between participants whether they were exposed to a noisy or non-noisy environment. In the non-noisy environment condition, filler sentences were the same as in Experiment 1; the non-noisy environment condition of Experiment 2 constituted a direct replication of Experiment 1. In the noisy environment condition, half the filler sentences were corrupted, making them ungrammatical. We intended to produce ten sentences with a verb that agreed with the NP in accusative case, 11 sentences with an NP with a case marker that made the sentence ungrammatical, and 11 sentences with a redundant or missing function word. Due to experimenter error, one sentence was corrupted with the two latter distortions, resulting in one sentence with two corruptions and only 31 ungrammatical filler sentences, instead of 32. The full set of stimuli is available on the OSF page.

**Participants**

We recruited participants on the crowd-sourcing platform Prolific for both experiments. In Experiment 1, we used Prolific to reach only participants who indicated that their first language is Russian. In Experiment 2, as per the preregistration, we reached participants who reported on Prolific that Russian is their first language and that they were born in one of the Ex-Soviet countries. Our reasoning was that, due to the large size of the Russophone diaspora and their highly variable command of the Russian language, participants born in one of the Ex-Soviet countries were more likely to grow up in a Russian-speaking environment, making them more

likely to be native in Russian. In both Experiments, participants were paid $3.00 for their participation. Exclusion criteria were identical in both experiments. In Experiment 1, we initially recruited 44 participants and excluded one participant for understanding less than 75% of grammatical filler sentences literally, two participants for reporting being born outside of the Ex-Soviet countries, and two participants for reporting that their first language is not Russian. The final sample size for Experiment 1 was 39. In Experiment 2, we initially recruited 260 participants, 130 per each of the two conditions of environment noise. We excluded eight participants for understanding less than 75% of the grammatical filler sentences literally, three for reporting being born outside the Ex-Soviet countries, and eight for reporting that their first language is not Russian. The final sample size for Experiment 2 was 241 participants, 124 in the non-noisy environment condition and 117 in the noisy environment condition.

**Procedure**

Participants were redirected from Prolific to the survey platform Qualtrics. All parts of the survey were in Russian. Participants read a consent form, provided their Prolific ID, and reported their country of birth, country of residence, first language, when and whether their second language was acquired, their gender, and their age. Participants were informed that they would be compensated the same regardless of their responses on these questions. Next, participants were instructed to reply with either yes or no to the items in the survey in one sitting. Each participant was randomly assigned (by Qualtrics) to one of the eight lists, with the constraint of keeping the lists with the same number of participants. For each participant, the order of the items was randomized. Upon finishing the survey, participants were automatically redirected back to Prolific, where they were compensated.

**Data Analysis**

All data were analyzed in R using the tidyverse libraries for data processing and visualization (R Core Team, 2019; Wickham et al., 2019), with Bayesian analyses conducted using the brms package (Bürkner et al., 2021). In all the models, we used trial-level literal interpretation (coded as 1 = literal, 0 = non-literal) as the binary dependent variable. Edit types were treatment coded with insertion edits as reference. Canonicality was sum coded (0.5 = canonical, -0.5 = non-canonical), as were environment noise (0.5 = noisy, -0.5 = non-noisy) and gender order in substitution trials (0.5 = plural first, -0.5 = female first). For each experiment, we fit 3 types of Models. Model 1 included the entire dataset and evaluated the effect of canonicality[4]. Model 2 included only the non-canonical data (following Gibson et al., 2013) and evaluated the effect of edit type. Model 3 included non-canonical substitution trials and compared the two types of substitution: pf and fp (**Table 1**). In experiment 2, all the models also evaluated the main effect of environment noise and its interaction with the other predictor in each model (canonicality, edit type, and substitution type, respectively). All the models used the maximal random effects structure justified by the design (i.e., the fixed effects specification was used for the random slopes within participants and items). We fit all models as a Bayesian logistic regression with the default priors in the brms package (i.e., flat priors for fixed effects). All model posteriors were sampled for 4000 iterations in total, 1000 of which were discarded as

[4] NB: Only Model 1 investigated the effect of canonicality; we do not explicitly test the interaction between canonicality and edit type for several reasons: 1) Literal interpretations of canonical sentences are at ceiling so there is insufficient variability to estimate differences between the types of canonical sentences; 2) The most likely edits that could have resulted in the canonical sentences are not the same as the ones for non-canonical versions of the same item so this is not a classic 2 (canonical, non-canonical) x 3 (deletions, insertions, substitutions) design in that sense; 3) Previous work (e.g., Gibson et al., 2013) did not analyze differences between canonical/plausible items. Therefore, we follow our preregistration in not investigating the rates of literal interpretation across canonical constructions.

warmup, with the exception of models with canonicality as the only predictor, which were fit with 2000 iterations, 500 of which were discarded warmup. If this specification resulted in any divergent transition or $\hat{R}$ values greater than 1.00, we adapted delta (the step size) and tree depth until there were no more divergent transitions.

## Results

### Experiment 1

The rates of literal interpretation across edit conditions are summarized in Figure 1 and Table 2. To study the effect of canonicality, we fit Model 1 as described above, with canonicality as a predictor, and random intercepts for participants and items. Participants were more likely to interpret canonical sentences literally than non-canonical sentences (Estimate = 3.72, 95% Credible Interval [CrI] = [3.08, 4.44]). To study the effect of edit type (deletion, insertion, substitution) on literal interpretations we fit Model 2 as described above with random intercepts for participants and items and random slopes for edit type with participants and items. Participants were less likely to interpret non-canonical sentences literally when they could have resulted from deletions than from insertions (Estimate = -0.91, 95% CrI = [-1.78, -0.02]). Participants were more likely to interpret non-canonical sentences literally when they could have resulted from substitutions than from insertions (Estimate = 0.70, 95% CrI = [-0.08, 1.44]).

We compared the two types of substitution edits to each other. There are two types of substitution edits because two gender orders could result in a substitution edit (female-plural and plural-female; see **Table 1**). Since we classify any item with a female NP and a plural NP as substitution edits regardless of order, we expected no difference in the proportion of literal interpretations between the two orders of NPs. To test this, we fit Model 3 as described above,

investigating the effect of gender order on literal interpretation, with random intercepts for participants and items, and random slopes for gender order within participants and items. The rates of literal interpretation were similar for non-canonical sentences that started with a female NP and those that started with a plural NP (Estimate = -0.26, 95% CrI = [-1.23, 0.62]).

Fillers items, which involved names that can be marked for case, were interpreted literally nearly all the time. Canonical fillers (16 out of 64 fillers) were interpreted literally 96.8% of the time. The rest of the fillers (32 out of 64 fillers) were general SVO sentences (not simply NP V NP), and they were interpreted literally 98.6% of the time.

**Table 2. Mean literal interpretation for all non-canonical conditions in Experiment 1.**

| Edit Type | Proportion of Literal Interpretations in Non-Canonical Conditions | Proportion of Literal Interpretations in Canonical Conditions |
|---|---|---|
| Deletion | 0.558 | 0.981 |
| Insertion | 0.686 | 0.982 |
| Substitution | 0.788 | 0.974 |

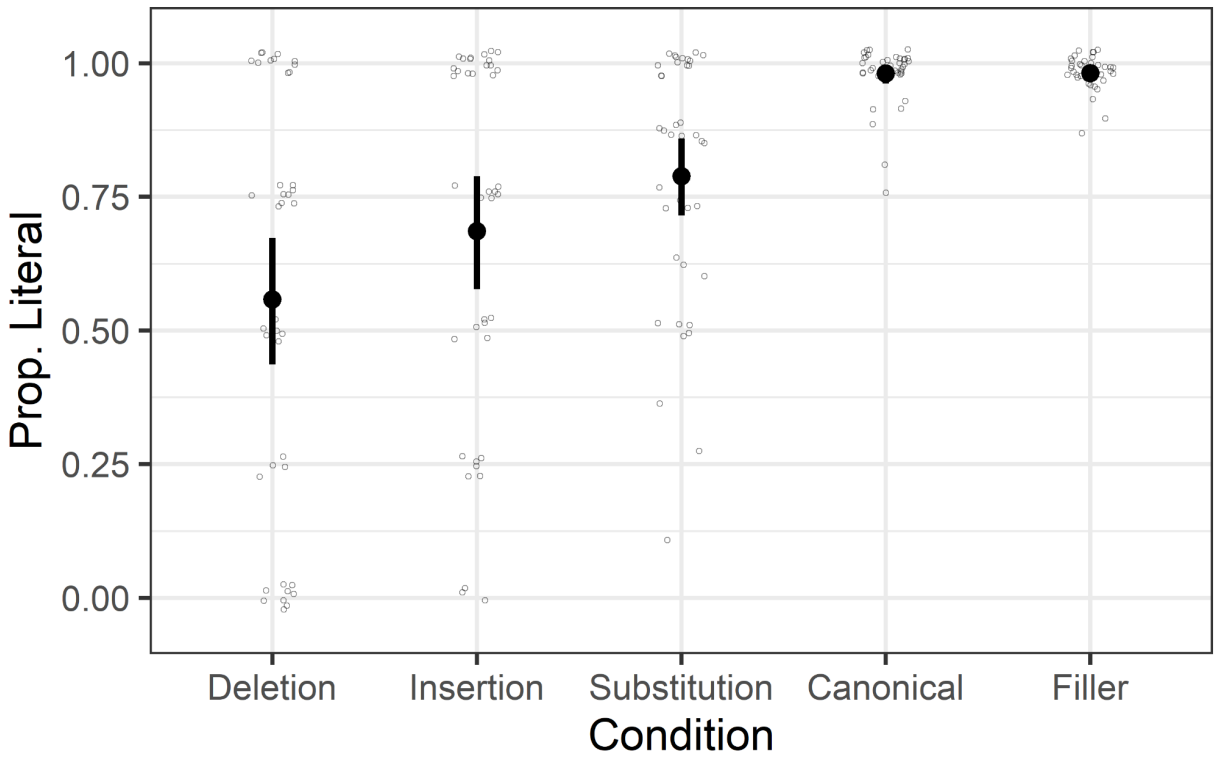


Figure 1. Results of Experiment 1. The proportion of literal interpretations for critical, non-canonical sentences with deletion, insertion, and substitution edits, in addition to canonical and filler sentences. Error bars are bootstrapped 95% confidence intervals. Unfilled points represent participant means.

**Experiment 2**

The rates of literal interpretation across environment noise, canonicality, and edit type conditions are summarized in Figure 2 and Table 3.

***Preregistered Analyses***

The experiment was preregistered with Bayesian logistic regression analyses with the default priors in the brms package (flat priors for fixed effects). To investigate the effect of

canonicality and environment noise on literal interpretations we fit Model 1 as described above, setting canonicality, environment noise, and their interaction as fixed effects with random intercepts for participants and items and a random slope for environment noise within verbs. Participants were more likely to interpret canonical sentences literally than non-canonical sentences (Estimate = 3.88, 95% CrI = [3.58, 4.20]), and they were less likely to interpret sentences literally when they were in a noisy environment (Estimate = -1.16 (95% CrI = [-1.67, -0.67]). Canonicality and environment noise do not appear to interact (Estimate = -0.37, 95% CrI = [-1.01, 0.24]). When investigating the effects of edit type (deletion, insertion, substitution) and environment noise on literal response using the non-canonical sentences, we fit Model 2 as described above with edit type, environment noise, and their interaction as fixed effects. We also added random intercepts for participants and items, random slopes for edit type within participants and items, and random slopes for environment noise and its interaction with edit type within items. We observed that participants were less likely to make a literal interpretation when reading sentences resulting from deletions than those resulting from insertions (Estimate = -0.45, 95% CrI = [-0.79, -0.10]), as in Experiment 1, and more likely to make a literal interpretation when reading sentences resulting from substitutions than those resulting from insertions (Estimate = 1.23, 95% CrI = [0.91, 1.57]), as in Experiment 1. Environment noise decreased the probability of choosing a literal interpretation (Estimate = -1.07, 95% CrI = [-1.64, -0.52]). The effects of edit types did not appear to be modulated by environment noise (deletion-environment noise interaction estimate = -0.05, CrI = [-0.62, 0.49]; substitution-environment noise interaction estimate = -0.12, CrI = [-0.70, 0.46]). The posterior distributions for the effects of edit type and environment noise are represented in **Figure 3**.

**Table 3. Mean literal interpretation for all non-canonical conditions in Experiment 2.**

| Environment Noise | Edit Type | Proportion of Literal Interpretations in Non-Canonical Conditions | Proportion of Literal Interpretations in Canonical Conditions |
|---|---|---|---|
| Non-Noisy | Deletion | 0.659 | 0.986 |
| Non-Noisy | Insertion | 0.736 | 0.988 |
| Non-Noisy | Substitution | 0.854 | 0.995 |
| Noisy | Deletion | 0.497 | 0.987 |
| Noisy | Insertion | 0.574 | 0.966 |
| Noisy | Substitution | 0.729 | 0.970 |

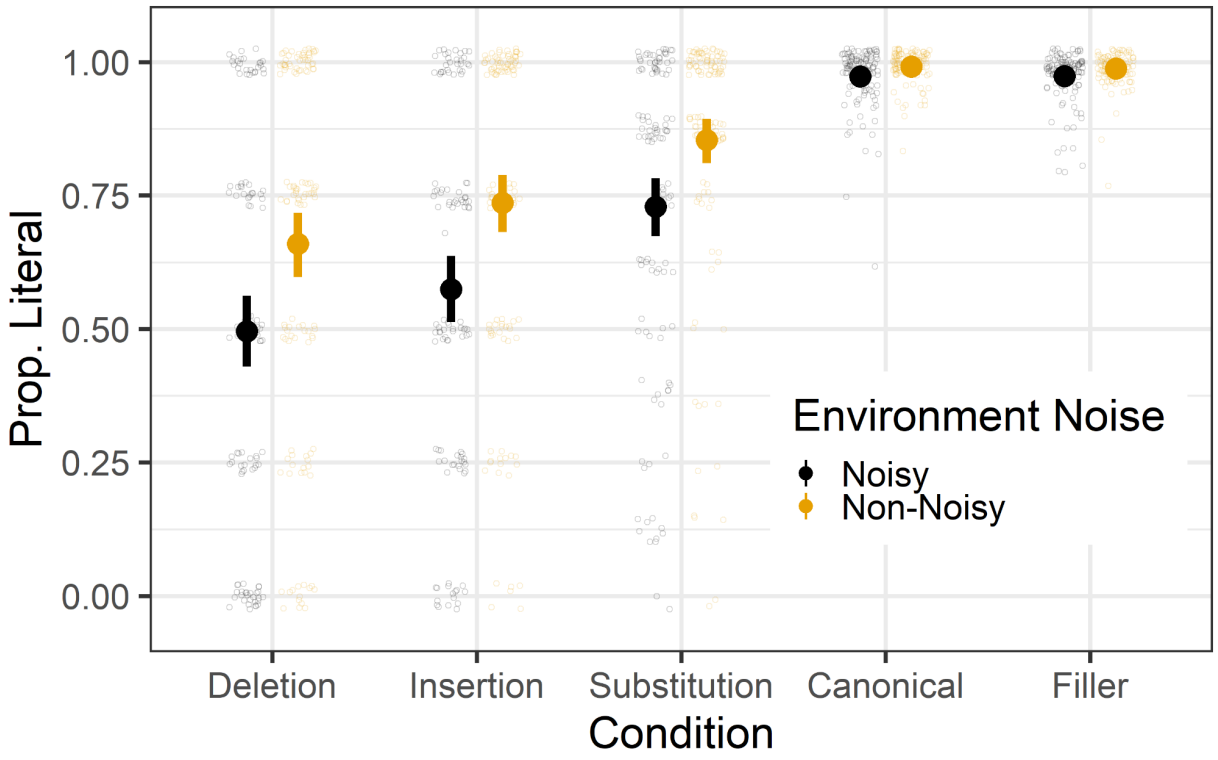


Figure 2. Results of Experiment 2. The proportion of literal interpretations for critical, non-canonical sentences with deletion, insertion, and substitution edits, in addition to canonical sentences and filler sentences. Participants in the noisy condition (where half of the filler

sentences were ungrammatical) are represented in black and participants in the non-noisy condition (all filler sentences were grammatical) are represented in orange. Error bars are bootstrapped 95% confidence intervals. Unfilled circles represent participant means.

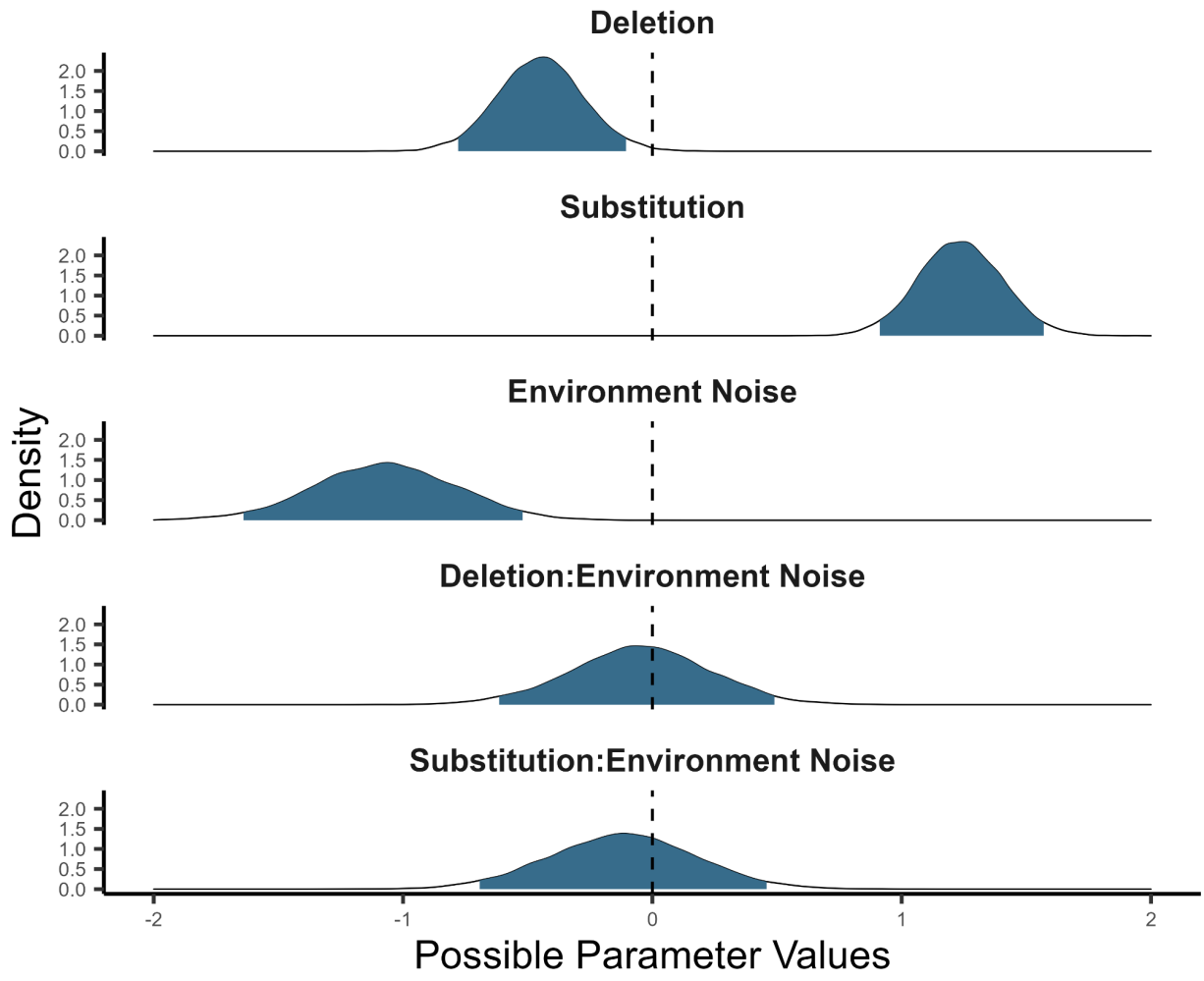


Figure 3. The posterior distributions from the model of environment noise and edit type. The shaded regions indicate 95% credible intervals.

***Exploratory Analyses***

As in Experiment 1, we tested whether the order of NPs within the substitution condition (female-plural or plural-female gender order) affected participants' literal interpretations. We did

not expect to find an effect of gender order. For this purpose, we analyzed the non-canonical substitution trials only, following Model 3 as described above, setting literal interpretation as the binary dependent variable and gender order (plural NP first or plural NP last) and environment noise as predictors (**Figure 4**). We set random intercepts for participants and items, with random slopes for gender order within participants and items, and random slopes for environment noise and its interaction with gender order within items. Unlike in Experiment 1, items that started with a plural NP were less likely to be interpreted literally than that started with a female NP (Estimate = -1.45, 95% CrI = [-2.27, -0.75]). While the effect of environment noise is similar to the previous analyses (a noisy environment increased the rate of non-literal interpretations; Estimate = -1.35, CrI = [-2.07, -0.68]), the interaction effect present between gender order and environment noise (Estimate = 1.02, 95% CrI = [0.30, 1.78]) suggests that the effect of gender order was smaller in the noisy environment than in the non-noisy environment.

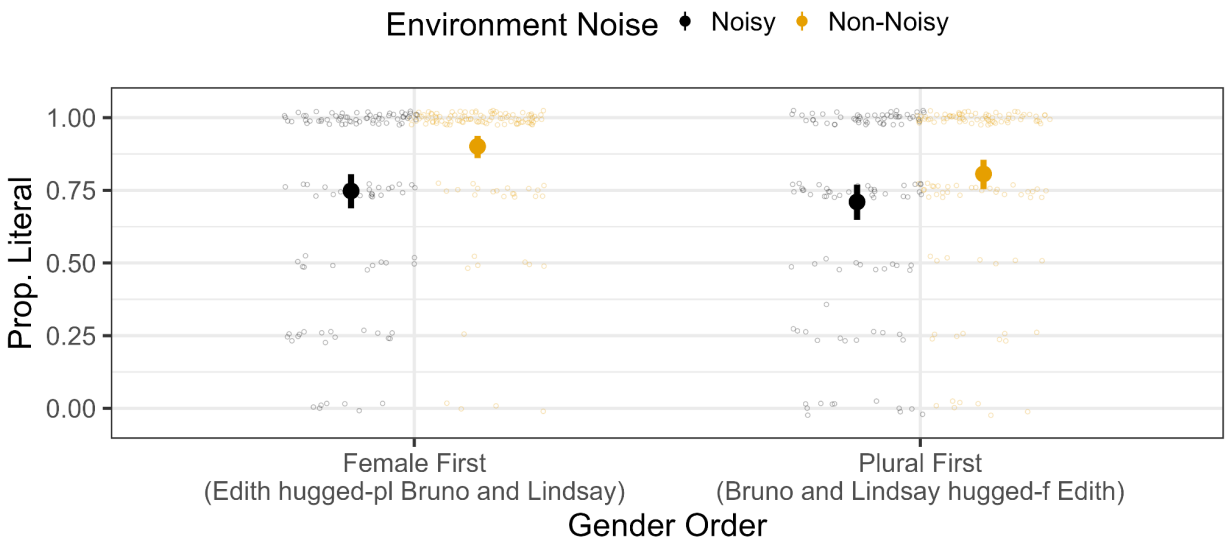


Figure 4. The proportion of literal interpretations for the two types of non-canonical substitution edits" female-first and plural-first. Participants in the noisy condition (where half of the filler sentences were ungrammatical) are represented in black and participants in the non-noisy

condition (all filler sentences were grammatical) are represented in orange. Error bars are bootstrapped 95% confidence intervals. Unfilled circles represent participant means.

Since the two substitution conditions substantially differed from each other, we tested the robustness of our results by fitting Model 2 again while only including substitution items that started with a plural NP, thus decreasing the effect size of substitution edits compared to insertion edits. All other model specifications remained the same. The results were consistent with the analysis including all the substitution data: participants were more likely to interpret literally sentences with substitution edits than sentences with insertion edits (Estimate = 0.93, 95% CrI = [0.56, 1.33]).

Fillers items, which involved names that can be marked for case, were interpreted literally nearly all the time in both the noisy and non-noisy environments. In the non-noisy environment, where all fillers were grammatical, canonical fillers (16 out of 64 fillers) were interpreted literally 99.2% of the time, while non-canonical fillers (16 out of 64 fillers) were interpreted literally 97.7% of the time. The rest of the fillers (32 out of 64 fillers) were general SVO sentences (not simply NP V NP), and they were interpreted literally 99.0% of the time. In a noisy environment, where half the filler items were ungrammatical (evenly distributed among filler types), canonical fillers (16 out of 64 fillers) were interpreted literally 98.5% of the time, non-canonical fillers (16 out of 64 fillers) were interpreted literally 95.0% of the time, and the general fillers (32 out of 64 fillers) were interpreted literally 97.8% of the time.

## Discussion

The current study showed that participants may interpret perfectly plausible and unambiguous sentences non-literally if they are unlikely under the structural prior. Previous

work showed that participants may interpret sentences non-literally if they are implausible (i.e., the sentences are unlikely under the meaning prior; Gibson et al., 2013; Zhan et al., 2017) or if they have an exceedingly rare structure (Keshev & Meltzer-Asscher, 2021). For example, topicalized sentences with OSV word order were studied in English and Chinese, where OSV word order has a frequency of 0.001 and 0.015, respectively (Liu et al., 2020). In this paper, we explored Russian in the noisy channel framework—a first, to our knowledge—using simple and short sentences while manipulating the verb agreement suffix to create SVO or OVS sentences. In Russian, SVO sentences are more frequent than OVS sentences, but, unlike non-canonical constructions in previous work on the structural prior (e.g., Liu et al., 2020), OVS word order is still used regularly in conversation and is not as rare (84.18% and 8.99%, respectively, in morphologically unambiguous sentences, and 87.15% and 7.58%, respectively, in ambiguous sentences; Berdicevskis & Piperski, 2020). Canonical (SVO) sentences were interpreted literally more often than non-canonical sentences (OVS). Moreover, non-canonical sentences had different proportions of literal interpretations, depending on the type of the underlying edit. Deletion edits were interpreted non-literally most often, followed by insertions and then substitutions. Conducting the experiment in Russian allowed for minimal pair comparisons, such that only one character edit separated any two conditions. Additionally, in Experiment 2, we manipulated environment noise between participants by making half the filler sentences ungrammatical for participants in the noisy environment condition. Participants in the noisy environment condition endorsed more non-literal interpretations of non-canonical sentences across the board than their counterparts in the non-noisy environment condition. In sum, manipulating the structural prior probability of the stimuli results in similar behavior as manipulating the semantic prior probability of the sentences. Like implausible sentences, non-

canonical sentences were often interpreted non-literally, with frequency that was dependent on the type of underlying edit and the environment noise.

In Experiment 2, we found that the different types of non-canonical substitution edits (e.g., female-first "Edith hugged-pl Bruno and Lindsay" and plural-first "Bruno and Lindsay hugged-f Edith") were not equally likely to be interpreted literally. Specifically, the plural-first substitution edits were more likely to be interpreted literally than female-first substitution edits, and that this effect was smaller in the noisy environment than in the non-noisy environment (**Figure 4**). This effect was not detected in Experiment 1, but, descriptively, the data in Experiment 1 pointed in the same direction. According to the noisy-channel framework, gender order should not affect the rate of non-literal interpretation as long as the noise likelihood is the same. Therefore, it was unexpected that, within the substitution edit type, sentences that started with a female NP were interpreted literally more often than sentences that started with a plural NP. One potential explanation for this result is that the plural NPs in the stimuli always consisted of a male name followed by a female name (e.g., "Joe and Rachel"). Therefore, in sentences where a plural NP was followed by a verb with a singular feminine agreement suffix (singular masculine is not used in the substitution condition because it is unmarked), the proximity of the verb to the female name in the plural NP may have created some 'local coherence' effects (Tabor et al., 2004) or closest conjunct agreement (Willer Gold et al., 2018). This resulted in participants more often failing to notice the discrepancy between the plural subject and the singular feminine verb suffix. In contrast, when a singular female NP is immediately followed by a plural verb, the mismatch is particularly noticeable, thus increasing the likelihood of a literal interpretation. Future work is needed to test this post-hoc explanation.

This work also sheds light on the noise model that comprehenders use when making inferences. Previous work has shown that comprehenders may assume edits to entire words, such as word deletions, insertions, and exchanges (Gibson et al., 2013; Poppels & Levy, 2016) and that deletions appear to be more likely than insertions, which appear to be more likely than exchanges. In the current study, we show that participants reason about edits to letters within a word, specifically bound morphemes, in a similar way to how they reason about words. Furthermore, the results lend support to the idea that Levenshtein distance provides a useful approximation of the edit likelihoods at both levels of granularity. Previous work (Ryskin et al., 2021) indicated that orthographic/phonetic distance within a single word was related to the likelihood of noisy-channel inference, but the relative probabilities of error types were not systematically explored. Here, the stimuli always differed only by one letter from their more plausible alternative, revealing that deletions are more probable than insertions, which are in turn more probable than substitutions, under the noise model used by readers in this experiment. This pattern is analogous to that observed when edits involve whole words and is consistent with the assumption that a substitution reflects both a deletion and an insertion. Other noise models could, in principle, have been possible. For instance, one might imagine a noise model under which any change to the total length of the string is less likely. In that case, a substitution could be viewed as the most probable error because it constitutes one edit that maintains the correct number of letters in a word (in contrast to deletions or insertions which change the number of letters). However, such a noise model is not supported by the data in the current experiment. Further investigation of the representations in the noise model, and how they may operate at different levels of granularity, will be important for sharpening the predictions of the noisy-channel framework.

Our findings address a previously raised critique of the noisy channel processing framework. Past studies (e.g., Gibson et al., 2013; Poppels & Levy, 2016) used implausible stimuli, like "the mother gave the daughter to the candle" or "the package fell from the floor to the table." It is possible that what drove the experimental results was, in part, participants' effort to deal with implausible sentences in an experimental setting. That is, the effects reflected processing of unnatural stimuli, not noisy channel processing in everyday communication. Task demands would not explain why sentences formed by some types of edits should be reliably more likely to be interpreted non-literally than others, but their effects on sentence interpretation are important to understand, nonetheless. In the present study, we used simple plausible sentences, and these gave rise to the same patterns of results as the previous studies. Therefore, the patterns that emerge in experiments on noisy-channel processing are unlikely to be a side-effect of implausible sentences in an experimental context. Rather, this work suggests that noisy-channel inferences are part and parcel of everyday human sentence comprehension across languages. More generally, we provide further evidence that many aspects of human cognition involve rational inference under uncertainty.

**Acknowledgements**

We would like to thank the three anonymous reviewers, and the audience at the 2022 Conference on Human Language Processing for their helpful comments and suggestions. We also thank the National Science Foundation for their grant support.

**Funding Statement**

The work was supported by a grant from the National Science Foundation (Award 2121074), "CompCog: Noisy-channel processing in human language understanding" (PI Gibson).

## References


Berdicevskis, A. & Piperski, A. (2020). Corpus evidence for word order freezing in Russian and German. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 26-33, Barcelona, Spain (Online). Association for Computational Linguistics.

Bouma, G. (2011). Production and comprehension in context. In A. Benz & J. Mattausch (Eds.), *Bidirectional optimality theory* (pp. 169–189). John Benjamins Publishing Company.

Bürkner, P.-C., Gabry, J., Weber, S., Johnson, A., & Modrak, M. (2021). *brms: Bayesian Regression Models using "Stan"* (2.15.0). https://CRAN.R-project.org/package=brms

Clayards, M., Tanenhaus, M. K., Aslin, R. N., Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues, *Cognition*, 108(3), 804-809. https://doi.org/10.1016/j.cognition.2008.04.004.

Cutter, M. G., Filik, R., & Paterson, K. B. (2022). Do readers maintain word-level uncertainty during reading? A pre-registered replication study. Journal of Memory and Language, 125, 104336.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–8056.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic

Inference. Trends in Cognitive Sciences, 20(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Jakobson, R. (1971). *Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutungen der russischen Kasus: Vol. II* (Originally published 1971, pp. 23–71). De Gruyter Mouton. https://doi.org/10.1515/9783110873269.23

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology*, *124*, 101359. https://doi.org/10.1016/j.cogpsych.2020.101359

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62 (1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Leckey, M., & Federmeier, K. D. (2019). The P3b and P600(s): Positive contributions to language comprehension. Psychophysiology, 0(0), e13351. https://doi.org/10.1111/psyp.13351

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Levy, R. (2008). A Noisy-Channel Model of Human Sentence Comprehension under Uncertain Input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. https://www.aclweb.org/anthology/D08-1025

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090. https://doi.org/10.1073/pnas.0907664106

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2020, September). *Structural frequency effects in comprehenders' noisy-channel inferences*. 26th Conference on Architectures and Mechanisms for Language Processing (AMLaP).

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., & Futrell, R. (2022). *Experimentally measuring the redundancy of grammatical cues in transitive clauses* (arXiv:2201.12911). arXiv. https://doi.org/10.48550/arXiv.2201.12911

Münte, T.F., Heinze, H.-J., Matzke, M., Wieringa, B.M., Johannes, S., 1998. Brain potentials and syntactic violations revisited: No evidence for specificity of the syntactic positive shift. Neuropsychologia 36 (3), 217–226. https://doi.org/ 10.1016/S0028-3932(97)00119-X.

Nathaniel, S., Ryskin, R., & Gibson, E. (2018, March). *The Effect of Context on Noisy-Channel Sentence Comprehension*. 31st Annual CUNY Conference on Human Sentence Processing.

Osterhout, L., Holcomb, P.J., 1992. Event-related brain potentials elicited by syntactic anomaly. J. Mem. Lang. 31 (6), 785–806. https://doi.org/10.1016/0749-596X(92) 90039-Z.

Poliak, M., Ryskin, R., Braginsky, M., & Gibson, E. (2023, March 4). It's not What You Say but How You Say It: Evidence from Russian Shows Robust Effects of the Structural Prior on Noisy Channel Inferences. Retrieved from osf.io/8tygf

Poppels, T., & Levy, R. P. (2016). Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 378–383.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nature Human Behaviour, 2(9), 693. https://doi.org/10.1038/s41562-018-0406-4

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, 107855. https://doi.org/10.1016/j.neuropsychologia.2021.107855

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124-1131.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611. https://doi.org/10.1016/j.tics.2014.08.001

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Willer Gold, J., Arsenijević, B., Batinić, M., Becker, M., Čordalija, N., Kresić, M., Leko,

N., Marušič, F. L., Milićev, T., Milićević, N., Mitić, I., Peti-Stantić, A., Stanković, B., Šuligoj, T., Tušek, J., & Nevins, A. (2018). When linearity prevails over hierarchy in syntax. *Proceedings of the National Academy of Sciences*, *115*(3), 495–500. https://doi.org/10.1073/pnas.1712729115

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. Psychological Review, 114(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Zhan, M., Levy, R., & Gibson, E. (2017, March). *Rational inference and sentence interpretation in Mandarin Chinese*. 30th CUNY Conference on Human Sentence Processing.

## Appendices

### Appendix A

We conducted a post-hoc study to estimate how clearly names were feminine or masculine. The study was conducted in Russian on the Qualtrics platform with 20 participants from Prolific. Participants self-identified their L1 as Russian, except for one participant who self-identified their L1 as Ukrainian. We chose to include the L1 Ukrainian speaker participant in the data because of the similarity between Ukrainian and Russian language and culture. In the study, all participants saw the 16 names in random order and were asked to rate how likely the name is to belong to a male or a female on a scale of 0 (definitely male) to 10 (definitely female). Mean ratings per name are reported in **Table 4** and individual participant ratings are visualized in **Figure 5**. Overall, the name gender ratings were consistent with the intended categories.

**Table 4. Mean ratings of the names used in the study. 0 = Definitely Male, 10 = Definitely Female.**

| Adele (Адель) | Jane (Джейн) | Jacqueline (Жаклин) | Kate (Кейт) | Lindsey (Линдси) | Rachel (Рейчел) | Scarlett (Скарлет) | Edith (Эдит) |
|---|---|---|---|---|---|---|---|
| 9.9 | 9.85 | 9.65 | 9.35 | 9.5 | 9.9 | 9.85 | 9.15 |

| Bruno (Бруно) | Joe (Джо) | Leo (Лео) | Matteo (Матео) | Romeo (Ромео) | Teo (Тео) | François (Франсуа) | Charlie (Чарли) |
|---|---|---|---|---|---|---|---|
| 0.64 | 1.4 | 0.65 | 0.55 | 0.3 | 1.25 | 1.65 | 2.05 |

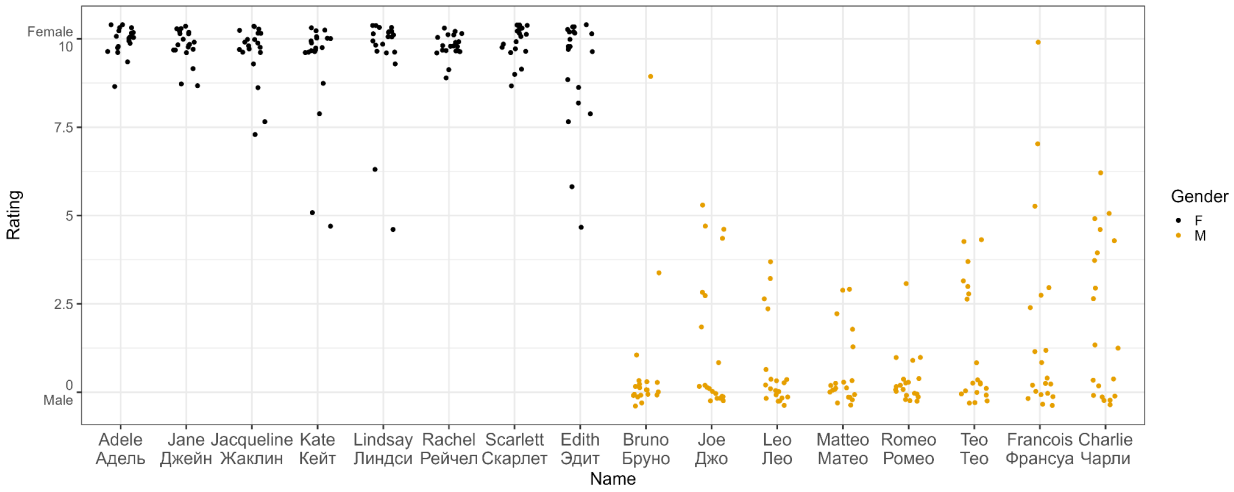


**Figure 5**. Participants rated how likely a name is to belong to a man or to a woman (definitely man = 0, definitely woman = 10). Each point is an individual response.

**Appendix B**

For each list, for each item, we randomized whether "yes" or "no" indicate a literal response. The randomization process did not balance the number of literal "yes" and "no" responses, so, for some conditions, "yes" indicated a literal interpretation more frequently than for other conditions. Specifically, in Experiment 1, the proportion of "yes" as literal response in deletions, insertions, and substitutions was 49%, 33%, and 53%, respectively. In Experiment 2, the proportion of "yes" as literal response in deletions, insertions, and substitutions was 49%, 37%, and 53%, respectively. To make sure that our results are not contingent on this imbalance, we have refitted all the models for all experiments with whether the literal response was "yes" or "no" as a covariate (coded as -0.5 = "no=literal", 0.5 = "yes=literal"). All other aspects of the model specifications remained identical to the models reported in the main text . The new estimates are summarized in **Table 5.** The inclusion of this covariate did not appear to meaningfully affect the estimates of predictors of interest.

**Table 5. A summary of model output with and without which response indicates literal interpretation as a covariate.**

| Experiment | Model | Effect | Original Values | Refitted Values |
|---|---|---|---|---|
| 1 | 1 | Canonicality | 3.72, [3.08, 4.44] | 3.72, [3.10, 4.47] |
| 1 | 1 | Yes = Literal interpretation | NA | 0.00, [-0.41, 0.39] |
| 1 | 2 | Deletion | -0.91, [-1.78, -0.02] | -0.89, [-1.73, 0.00] |
| 1 | 2 | Substitution | 0.70, [-0.08, 1.44] | 0.73, [-0.05, 1.50] |
| 1 | 2 | Yes = Literal interpretation | NA | -0.13, [-0.65, 0.40] |
| 1 | 3 | Plural-first | -0.26, [-1.23, 0.62] | -0.24, [-1.21, 0.67] |

| | | | | |
|---|---|---|---|---|
| 1 | 3 | Yes = Literal interpretation | NA | 0.28, [-0.48, 1.09] |
| 2 | 1 | Canonicality | 3.88, [3.58, 4.20] | 3.87, [3.58, 4.19] |
| 2 | 1 | Environment Noise | -1.16, [-1.67, -0.67] | -1.15, [-1.66, -0.65] |
| 2 | 1 | Interaction of Canonicality and Environment Noise | -0.37, [-1.01, 0.24] | -0.36, [-1.00, 0.23] |
| 2 | 1 | Yes = Literal interpretation | NA | -0.01, [-0.18, 0.16] |
| 2 | 2 | Deletion | -0.45, [-0.79, -0.10] | -0.43, [-0.76, -0.10] |
| 2 | 2 | Substitution | 1.23, [0.91, 1.57] | 1.25, [0.92, 1.60] |
| 2 | 2 | Environment Noise | -1.07, [-1.64, -0.52] | -1.07, [-1.64, -0.52] |
| 2 | 2 | Interaction of Environment Noise and Deletion | -0.05, [-0.62, 0.49] | -0.04, [-0.59, 0.49] |
| 2 | 2 | Interaction of Environment Noise and Substitution | -0.12, [-0.70, 0.46] | -0.11, [-0.70, 0.45] |
| 2 | 2 | Yes = Literal interpretation | NA | -0.15, [-0.38, 0.09] |
| 2 | 3 | Plural-first | -1.45, [-2.27, -0.75] | -1.46, [-2.27, -0.78] |
| 2 | 3 | Environment Noise | -1.35, [-2.07, -0.68] | -1.37, [-2.08, -0.67] |
| 2 | 3 | Interaction between Gender Order and Environment First | 1.02, [0.30, 1.78] | 1.04, [0.32, 1.79] |
| 3 | 3 | Yes = Literal | NA | 0.12, [-0.29, 0.53] |

| | | | | |
|---|---|---|---|---|
| | | interpretation | | |
| 2 | 4 | Substitution | 0.93, [0.56, 1.33] | 0.96, [0.59, 1.38] |
| 2 | 4 | Yes = Literal interpretation | NA | -0.37, [-0.64, -0.09] |