Rational Inference Underlies Judgments of Grammatical Well-Formedness

Moshe Poliak* (MIT), Aixiu An* (MIT), Roger Levy (MIT), Edward Gibson (MIT)
*indicated equal contribution; corresponding email moshepol@mit.edu

Background: Acceptability judgments are the main tool for investigating the grammar of a language, both with human subjects[1-4] and with large language models[5]. But what makes a sentence more or less acceptable? In this project, we evaluate three potential mechanisms across 8 languages. (1) Ever since Chomsky[6], it has been a standard assumption that grammaticality exists on a spectrum. Partially formalizing this idea, Pullum[7] proposed a framework in which the grammar of a language is a set of binary constraints, and the more constraints a sentence violates, the less grammatical that sentence is (**Equation 1**). (2) Another potential mechanism builds on mechanism 1 but also considers sentence length, such that longer sentences are less acceptable[8] (**Equation 2**). Whereas mechanisms 1-2 evaluate grammatical well-formedness from a linguistic structural perspective, we propose a different mechanism rooted in rational communication. (3) If the goal of language is to successfully exchange information, then **grammatical well-formedness should reflect how easy it is to infer the speaker's intention**. Therefore, mechanism 3 predicts that the **higher the percentage of uncorrupted information** in a sentence, the more acceptable the sentence will be (**Equation 3**), in addition to longer sentences being less acceptable. We operationalize this by dividing the number of corruptions in a sentence by the sentence's length.

Method: We evaluated the 3 mechanisms above using 8 experiments with the same design across Danish, English, French, German, Hindi, Korean, Mandarin, and Russian (Ns= 40, 40, 40, 40, 33, 36, 30, 41 respectively, after exclusions). For each language, we selected 72 sentences of different lengths (range: [4,43], median: 15), creating 4 conditions from each sentence: original, 1 transposition, 3 transpositions, and a shuffled word order (see **Table 1**). Participants were presented with all the sentences once, with semi-random assignment of condition to sentence such that each participant saw each condition the same number of times. Participants were asked to rate how natural each sentence is and then responded to a comprehension question about the sentence (inclusion criterion: >80% accuracy).

Results: The results from all languages are represented in **Figure 1**. We fit 3 cumulative Bayesian regressions with random intercepts for participants and for items within languages, varying the fixed effects according to **Equations 1-3**, and compared their predictive abilities using WAIC[9-11]. **Equation 3** had the best predictive ability by far, followed by **Equation 2** (ELPD Difference from Equation 3 = -1011, SD = 46.1), which was not substantially better than **Equation 1** (ELPD Difference from Equation 3 = -1045.9, SD = 48.7). Moreover, this inferential finding replicated within each language separately, as is also seen in the descriptive **Figure 1**, where sentences with 1-5 corruptions increase in acceptability the longer they are, for all languages.

Discussion: We find that the best explanation for the grammaticality of sentences is rooted in rational comprehension: the grammaticality of sentences reflects how easy it is to recover what the speaker intended, adding to the growing evidence that the goal of language comprehension is to understand the message that the speaker intended to communicate.

Equation 1: $grammaticality = \beta_0 + \beta_1 * corruptions$

Equation 2: $grammaticality = \beta_0 + \beta_1 * corruptions + \beta_2 * |s|$

Equation 3: $grammaticality = \beta_0 + \beta_1 * |s| + \beta_2 * \frac{corruptions}{|s|}$

Table 1. A sample item in English.

condition	sentence
original	A ball flying in the air can hurt.
1 transposition	A ball flying in air the can hurt.
3 transpositions	A ball in flying can the air hurt.
shuffled word order	In flying can ball a hurt air the.

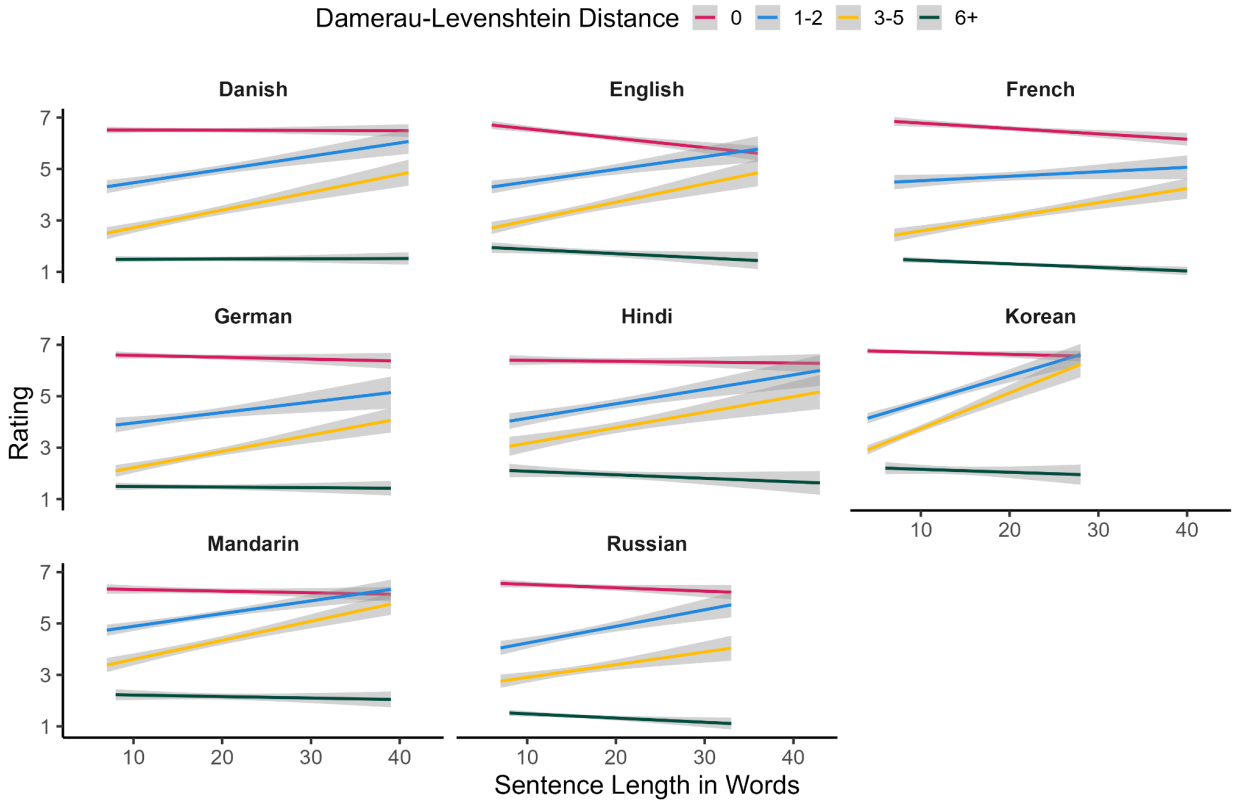

Figure 1. The line of best fit for acceptability rating as predicted by sentence length and the Damerau-Levenstein distance between the original and corrupted sentences, split by language. The distance is the minimal number of words that need to be deleted, inserted, substituted, or transposed with the neighboring word to arrive from the presented sentence to the original sentence that was collected from the UD treebank.

References: [1]Schütze, 1996; [2]Cowart, 1997; [3]Myers, 2009; [4]Sprouse et al., 2013; [5] Warstadt et al., 2019; [6]Chomsky (1964); [7]Pullum (2020); [8]Lau et al. (2017); [9]R core team (2024); [10]Wickham et al. (2024); [11]Bürkner (2017).