Incorporating Formulaicness in the Automatic Evaluation of Naturalness: A Case Study in Logic-to-Text Generation

Eduardo Calò[1] **Guanyi Chen**[2] **Elias Stengel-Eskin**[3] **Albert Gatt**[1] **Kees van Deemter**[1]
[1]Utrecht University [2]Central China Normal University [3]University of Texas at Austin
{e.calo, a.gatt, c.j.vandeemter}@uu.nl
g.chen@ccnu.edu.cn esteng@utexas.edu

Abstract

Data-to-text natural language generation (NLG) models may produce outputs that closely mirror the structure of their input. We introduce *formulaicness* as a measure of the output-to-input structural resemblance, proposing it as an enhancement for reference-less naturalness evaluation. Focusing on logic-to-text generation, we construct a dataset and train a regressor to predict formulaicness scores. We collect human judgments on naturalness and examine how incorporating formulaicness into existing metrics affects alignment with these judgments.

1 Introduction

When generative models are provided with structured input data (e.g., logical formulae, tables, RDF triples, etc.) and are tasked with generating textual descriptions of this input data, they sometimes produce outputs that mirror the structure of the input very closely. This can be undesirable in domains and applications that require fluent output, instead of the stilted texts exemplified in Table 1. On the other hand, such structure-preserving renderings of inputs might be desirable, depending on the application domain and/or target audience. For instance, in contexts like teaching logic connectives, where precise and formulaic mappings may be preferred, one may want to favor formulaic output such as that in the top row of Table 1.

These considerations have implications for evaluating the broader notion of *naturalness* in text generation. One important but previously unexplored feature is what we term **formulaicness**: *the degree to which a data-to-text NLG system's output explicitly preserves the structural form of its input*. High formulaicness reflects a close, template-like correspondence to the input, while low formulaicness indicates greater abstraction or paraphrasing. As shown in Table 1, excessive formulaicness can lead to stilted outputs, even in texts generated by large language models (LLMs). For instance, a less formulaic realization of the logic-to-text input might be: *There is exactly one large cube, and there is no dodecahedron behind it*.

In this paper, we propose a methodology for measuring formulaicness and investigate whether incorporating this measure into other reference-less (Ito et al., 2025) evaluation metrics (including LLM-based ones) can improve the automatic assessment of naturalness. Such metrics typically capture features like fluency, grammaticality, or readability (e.g., Kann et al., 2018; Groves et al., 2018; Çano and Bojar, 2020; Zhu and Bhat, 2020; Liu et al., 2021; Nguyen et al., 2024), but are not always consistent with human judgments (Novikova et al., 2017). Our main question is: *Does incorporating formulaicness in evaluation metrics improve the automatic assessment of naturalness?*

We focus on logic-to-text generation, a task with a long tradition in NLG (e.g., Wang, 1980) that has drawn significant renewed interest more recently (e.g., Haroutunian et al., 2023; Wu et al., 2023), in part due to the growing attention to the reasoning abilities of LLMs (Cheng et al., 2025), with benchmarks and experimental work carried out, also regarding first-order logic (e.g., Tian et al., 2021; Han et al., 2024; Karia et al., 2024). Logic-to-text generation can be particularly valuable when logical formulae are difficult to interpret, for instance for beginners (Rector et al., 2004), since translating them into natural language can enhance comprehension. For example, Mpagouli and Hatzilygeroudis (2009) employed a logic-to-text system to translate first-order logic formulae into English in a classroom setting.

2 Methods

We constructed a dataset consisting of texts (**a.**) derived from logical formulae paired with formulaicness scores in the [0, 1] range (**b.**). We used this

Proceedings of the 18th International Natural Language Generation Conference, pages 352–365
October 29–November 2, 2025.

<table>
<tr><th>Subtask</th><th>Dataset</th><th>Input</th><th>Output</th><th>Model</th></tr>
<tr><td>Logic-to-Text</td><td>Grade Grinder Corpus (Barker-Plummer et al., 2011)</td><td>$\exists x \forall y \forall z$ (
$\text{Cube}(x) \wedge \text{Large}(x) \wedge$
(
$(\text{Cube}(y) \wedge \text{Large}(y) \wedge \text{Dodec}(z)) \rightarrow$
$(x = y \wedge \neg \text{BackOf}(z, y))$
)
)</td><td>There exists a cube x that is large, and for all y and z, if y is a cube and large and z is a dodecahedron, then x is y and z is not behind y.</td><td>Qwen3-32B (Yang et al., 2025)</td></tr>
<tr><td>Table-to-Text</td><td>numericNLG (Suadaa et al., 2021)</td><td><table><tr><th>Genre</th><th>Sentences</th><th>Length</th><th>Yield</th><th>Precision</th></tr><tr><td>News*</td><td>100</td><td>19.3</td><td>142</td><td>78.9</td></tr><tr><td>News</td><td>100</td><td>19.3</td><td>144</td><td>70.8</td></tr><tr><td>Wiki</td><td>100</td><td>21.4</td><td>178</td><td>61.8</td></tr><tr><td>Web</td><td>100</td><td>19.2</td><td>165</td><td>49.1</td></tr><tr><td>Total</td><td>300</td><td>20.0</td><td>487</td><td>60.2</td></tr></table></td><td>The data provided shows the results of different genres of text, including News with and without a star, Wiki, and Web, in terms of the number of sentences, length, yield, and precision. The News with a star has one hundred sentences, an average length of nineteen point three words, a yield of one hundred forty two, and a precision of seventy eight point nine percent. [...]</td><td>Llama-3.3-70B-Instruct (Grattafiori et al., 2024)</td></tr>
<tr><td>RDF-to-Text</td><td>WebNLG (Gardent et al., 2017)</td><td>[
"San_Sebastián_de_los_Reyes | isPartOf | Community_of_Madrid",
"ENAIRE | city | Madrid",
"Adolfo_Suárez_Madrid-Barajas_Airport | location | San_Sebastián_de_los_Reyes",
"San_Sebastián_de_los_Reyes | country | Spain",
"Adolfo_Suárez_Madrid-Barajas_Airport | operatingOrganisation | ENAIRE"
]</td><td>San Sebastián de los Reyes is part of the Community of Madrid.
ENAIRE is located in the city of Madrid.
Adolfo Suárez Madrid–Barajas Airport is located in San Sebastián de los Reyes.
San Sebastián de los Reyes is in the country of Spain.
Adolfo Suárez Madrid–Barajas Airport is operated by ENAIRE.</td><td>DeepSeek-V3 (DeepSeek-AI et al., 2025)</td></tr>
</table>

Table 1: Examples of formulaic outputs from structured inputs of various kinds. The models were prompted "*Convert this input data into English: {Input} ONLY RETURN THE TEXT.*" via Hugging Face Playground (Wolf et al., 2020) with temperature set to 0.

data to train a regressor to predict formulaicness $F(t)$ for a given text t (**c.**). We collected human judgments of naturalness (**d.**) and explored the impact of incorporating formulaicness into other metrics ($M(t)$) (**e.**) using a weighted average:

$$\text{Combined}(t) = \frac{\alpha \cdot M(t) + \beta \cdot (1 - F(t))}{\alpha + \beta} \quad (1)$$

We used the complement $1 - F(t)$, so as to *minimize* formulaicness, e.g., in contexts where greater abstraction from the input is desirable. The weights α and β control the relative influence of M and F.

a. Texts We used a subset of the dataset from Calò et al. (2022), consisting of first-order logic formulae from the Grade Grinder Corpus (GGC; Barker-Plummer et al. 2011), paired with English translations generated using three systems, which generate texts of different degrees of formulaicness by design: (i) a system that generates literal translations of the formulae; (ii) Ranta (Ranta, 2011b), which performs syntactic optimizations to improve fluency; and (iii) LoLa, an extension of Ranta that applies logical optimizations to the input formula before verbalizing it (Appendix A for details). To complement the set, we included (iv) the human-written translations of the formulae; henceforth, we speak of Literal, Ranta, LoLa, and Human.

b. Scores In Calò et al. (2022), the three systems were ranked LoLa > Ranta > Literal in terms of fluency, based on human judgments, using TrueSkill (Herbrich et al., 2006; Sakaguchi et al., 2014), an algorithm for estimating relative performance scores, which returns a mean (μ) and a standard deviation (θ) for each system reflecting the final ranking (the higher the μ, the better the system). We heuristically interpret this fluency ranking as a proxy for formulaicness, with human-written texts considered the least formulaic overall.

To assign appropriate formulaicness scores to texts (i.e., ensuring Literal > Ranta > LoLa > Human on average formulaicness), we defined score bins for each system by leveraging the μ and θ values returned by TrueSkill.[1] The center of each bin was set to the corresponding TrueSkill μ (with higher μ indicating lower formulaicness; values were normalized to the [0, 1] range), and the bin width was set to $\pm\theta$ (clipped to be within [0, 1]). Within each bin, we randomly sampled floating-point values and took their complements ($1-x$) to synthesize a formulaicness score for each text.

The final dataset consists of 570 texts and corresponding formulaicness scores. We split the data into training, validation, and test sets using a 70-15-15 ratio, stratifying by formulaicness score and text length by token to ensure consistent distributions across splits. Table 2 presents a sample from the test set. See Appendix B for more details.

c. Predicting Formulaicness To model formulaicness, we fine-tuned a BERT-based (Devlin et al., 2019) regressor on the dataset described above, by adding a linear regression head on top of a pre-

[1] No TrueSkill values are available for human-written texts from Calò et al. (2022), since they were not evaluated in the original study. Following the earlier assumption that human texts are the least formulaic overall, we arbitrarily assigned Human a higher μ than LoLa with a low θ.

Text	Formulaicness
For all x, if x is a cube, then there is an element y such that y is a tetrahedron and x is in the same row as y and y is to the right of x.	0.85
A large cube is in front of a small cube.	0.10
a is a tetrahedron and e is a tetrahedron or a is a tetrahedron and f is a tetrahedron.	0.72
For all x, for all y, if y is large and y is a cube, then x is not to the right of y or x is small.	0.54

Table 2: Sample of texts from the test set with their associated formulaicness scores.

trained BERT model to predict a continuous score in the [0, 1] range. See Appendix C for details. We evaluated the final model on the held-out test set, obtaining a mean squared error (MSE) of 0.017 (R^2 = 0.813). See Appendix D for additional analyses.

d. Human Judgments To assess our metrics, we recruited 100 native English speakers via Prolific (median age = 38.5; 50% female), who were asked to rate the naturalness of each text in the held-out test set (N = 86; see above) on a 7-point Likert scale (Likert, 1932). Naturalness was defined following Howcroft et al. (2020) (see Appendix G for the full instructions). We used Qualtrics to set up the experiment. The 86 test texts were divided into four mutually exclusive batches (each containing ~20 texts). Each participant was assigned to one batch, so each text received ~25 annotations. The median completion time for the task was ~8 minutes. We included two attention checks to ensure data quality,[2] and randomized the order of texts to avoid order effects.

Inter-annotator agreement (IAA) was computed using Krippendorff's α (Krippendorff, 1980) and ranged from low to moderate across the four batches (0.19, 0.22, 0.43, 0.13), with an average of 0.25. Variation across batches can result from differences in item difficulty and annotator behavior, including variation in Likert-scale leniency. These IAA scores are consistent with prior findings that human judgment tasks in NLG often yield low agreement (van der Lee et al., 2021). Importantly, low IAA is not inherently problematic. It may reflect the subjective and challenging nature of the task, which is itself informative. For instance, Plank (2022) argues that human label variation is ubiquitous and that high IAA is typically achieved only under artificial conditions.

e. Baseline Metrics We used five reference-less metrics as approximations of the standardized notion of naturalness adopted in this study (details in Appendix E): GRUEN (Zhu and Bhat, 2020); Flesch Reading Ease score (FRE, Flesch, 1948); Perplexity (PPL, Jelinek et al., 1977); the Syntactic Log-Odds Ratio (SLOR, Pauls and Klein, 2012; Kann et al., 2018); Llama 3.1 8B Instruct as LLM-as-judge (LLAMA, Grattafiori et al., 2024). To make all metrics interpretable in a consistent way, where higher is better (= more natural), we inverted PPL and SLOR (henceforth, iPPL and iSLOR).

3 Results

For each text in the test set, we computed formulaicness, GRUEN, FRE, iPPL, iSLOR, LLAMA, and mean naturalness scores provided by participants. All baseline scores were normalized to the [0, 1] interval. For each baseline metric, we also computed combined score (Equation 1). Additional information is provided in Appendix F.

The Pearson correlation between predicted formulaicness scores and average human ratings is $r = -0.594$ ($p \ll 0.005$), indicating a highly significant negative correlation: more formulaic texts tend to be rated as less natural. Next, we evaluated the five baseline metrics, both in isolation and combined (Table 4). In isolation, GRUEN and FRE show statistically non-significant correlations; iPPL shows a moderate, statistically significant negative correlation; and iSLOR and LLAMA show a moderate, statistically significant positive correlation. When combined, all metrics yield moderate to strong, statistically significant positive correlations, with increases ranging from modest (iSLOR, LLAMA) to substantial (GRUEN, FRE, iPPL).

To assess the contribution of formulaicness when combined with baseline metrics, we conducted goodness-of-fit tests by fitting linear regression models to predict human naturalness ratings using: (i) each baseline metric alone, (ii) formulaicness alone, and (iii) their combinations (Table 5). The model using formulaicness alone explains a substantial portion of the variance in naturalness ratings (R^2 = 0.352), consistently outperforming models based solely on baseline metrics. Combining formulaicness with baseline metrics always improves model fit (R^2 Combined), yielding statistically significant gains across all metrics except

[2] All participants passed at least one of the attention checks.

Text	Naturalness	Formulaicness	Analysis
If it is not the case that c is a tetrahedron, then it is not the case that b is a tetrahedron.	5.08	0.918	High formulaicness (expected) due to repeated use of *it is not the case*; yet rated highly natural by humans, likely for its symmetrical structure.
There is an element x such that there is an element y such that there is an element z such that x is a cube, x is small, y is a cube, y is medium, z is a cube and z is large.	2.16	0.350	Low formulaicness (unexpected) despite low human-rated naturalness.
Some dodecahedron is not small.	4.40	0.040	Medium human naturalness (somewhat awkward phrasing); very low formulaicness, likely due to lack of literals or formulaic patterns.

Table 3: Examples illustrating divergences between human naturalness scores and formulaicness regressor scores.

Metric	r Baseline	r Combined	Δ
GRUEN	-0.156	0.547*	+0.703
FRE	0.108	0.542*	+0.434
iPPL	-0.580*	0.395*	**+0.975**
iSLOR	**0.526***	**0.624***	+0.098
LLAMA	0.391*	0.589*	+0.198

Table 4: Pearson correlations (r) with human naturalness ratings, for each metric alone (Baseline) and in combination with formulaicness (Combined). Δ indicates the gain from combining formulaicness with the baseline metric. Asterisks (*) indicate statistically significant ($p \ll 0.005$) values. **Boldfaced** is the higher value per column.

iPPL. The largest Δ is observed with FRE, while the highest overall R^2 is achieved when formulaicness is combined with iSLOR.

Metric	R^2 Baseline	R^2 Combined	Δ
Formulaicness	**0.352**	—	—
GRUEN	0.024	0.300*	+0.275
FRE	0.012	0.294*	**+0.282**
iPPL	0.336	0.156	−0.180
iSLOR	0.277	**0.390***	+0.113
LLAMA	0.153	0.346*	+0.193

Table 5: Performance of each metric in predicting human naturalness ratings. We report R^2 values for each metric alone (Baseline) and in combination with formulaicness (Combined). Δ indicates the gain from combining formulaicness with the baseline metric. Values with an asterisk (*) indicate statistically significant gains over the baseline metric alone ($p < 0.05$, t-test). Values underlined indicate the higher score between Baseline and Combined per metric. **Boldfaced** is the higher value per column.

To investigate discrepancies between human judgments of naturalness and formulaicness scores assigned by the BERT regressor, we performed a linguistic analysis on samples where human naturalness ratings diverged significantly from the BERT regressor predictions (see Table 3).

To examine how adding formulaicness affects baseline metrics, we selected examples with the largest deltas between baseline scores and scores combined with formulaicness (Table 6). GRUEN often overrates unnatural texts; combining it with formulaicness lowers the scores, improving alignment with human judgments. FRE is inconsistent, underestimating a natural text in one case and overestimating unnatural ones in others, with these errors reduced in the combined scores. iPPL shows the largest deltas, assigning extreme values that are moderated by the combination. iSLOR shows the lowest deltas, suggesting that it already correlates well with human ratings. LLAMA tends to give repetitive ratings, assigning high scores to both low and medium natural texts indiscriminately, mitigated by the combination.

4 Discussion and Conclusion

Formulaicness improves the evaluation of naturalness. Returning to our research question, the results in §3 suggest that formulaicness offers an effective enhancement to the automatic evaluation of naturalness in logic-to-text generation. It aligns well with human judgments and consistently improves the performance of baseline metrics in terms of correlation with human ratings (Table 4) and explanatory power in regression models (Table 5).

Formulaicness improves baseline metrics. The generally low values of the baseline metrics (Table 4 and Table 5) warrant further comment. Each metric captures different facets of naturalness (e.g., GRUEN targets grammaticality, FRE readability). The consistent improvements when adding formulaicness suggest that formulaicness captures an orthogonal dimension of naturalness, one not fully accounted for by grammaticality or readability alone. The one exception is SLOR, likely due to its theoretical properties: SLOR normalizes for unigram probabilities and sentence length (Appendix E), making it particularly effective in our task of logic-to-text generation, in which sentence length varies considerably, and logical variables and constants (e.g., x, y, which a language model treats as unigrams) appear regularly (e.g., Table 7).

Metric	Text	Baseline	Combined	Δ	Naturalness
GRUEN	*For all x, if it is not the case that x is a cube, then x is a tetrahedron and if it is not the case that x is a tetrahedron, then x is a cube.*	0.825	0.558	−0.266	3.54
	If b is a dodecahedron, then if it is not the case that b is in front of d, then it is not the case that b is in back of d.	0.811	0.551	−0.260	2.81
	If it is not the case that b is to the left of d and it is not the case that b is to the right of d, then b is a tetrahedron or d is a tetrahedron.	0.799	0.541	−0.258	3.76
FRE	*Only dodecahedra are larger than everything else.*	0.000	0.392	+0.392	4.69
	D is a cube, c is a cube and it is not the case that d or c is small.	0.837	0.559	−0.278	3.96
	For all x, if x is even, then it is not the case that x is a prime.	0.810	0.537	−0.273	4.32
iPPL	*No cube is large.*	0.000	0.487	+0.487	6.20
	For all x, if it is not the case that x is a cube, then x is a tetrahedron and if it is not the case that x is a tetrahedron, then x is a cube.	1.000	0.523	−0.477	3.54
	If it is not the case that b is to the left of d and it is not the case that b is to the right of d, then b is a tetrahedron or d is a tetrahedron.	0.958	0.501	−0.457	3.76
iSLOR	*No tetrahedron is the same size as any cube.*	0.282	0.482	+0.200	4.65
	There is an element x such that there is an element y such that there is an element z such that x is a cube, x is small, y is a cube, y is medium, z is a cube and z is large.	0.019	0.209	+0.190	2.16
	Only dodecahedra are larger than everything else.	0.352	0.536	+0.185	4.69
LLAMA	*There is an element y such that y is a dodecahedron and it is not the case that y is large.*	0.950	0.664	−0.286	3.96
	It is not the case that there is an element x such that there is an element y such that x is in back of y and it is not the case that x is larger than y.	0.950	0.669	−0.281	2.65
	It is not the case that some dodecahedron is large.	0.950	0.687	−0.263	4.88

Table 6: Samples with the top-3 largest Δs in score per metric between before (Baseline) and after (Combined) being combined with formulaicness vs. human judgments (Naturalness).

Generalizability beyond logic-to-text. As noted in §1, other NLG tasks such as table-to-text and RDF-to-text generation exhibit similar issues with formulaicness as logic-to-text. We hypothesize that our method could generalize to these tasks by helping identify formulaic outputs. For example, a less formulaic version of the RDF-to-text input in Table 1 might be: *San Sebastián de los Reyes is located in Spain, within the Community of Madrid. It is home to Adolfo Suárez Madrid–Barajas Airport, which is operated by ENAIRE, a company based in the city of Madrid.* The main challenge would be obtaining formulaicness scores comparable to those derived from Literal, Ranta, LoLa, and Human in §2. Evaluations of existing systems in these tasks (e.g., Nikiforovskaya and Gardent, 2024, for RDF-to-text) could serve as a starting point.

Conclusion. We introduced *formulaicness* as an enhancement for reference-less naturalness evaluation, focusing on logic-to-text generation. Our results show that incorporating formulaicness into existing metrics improves alignment with human judgments of naturalness.

Limitations

This study focused exclusively on logic-to-text generation into English as a case study. It remains to be seen whether our approach to modeling formulaicness generalizes effectively to other NLG tasks, such as table-to-text or RDF-to-text generation, and to other languages.

In our implementation, we intentionally avoided using LLMs for modeling formulaicness, and instead relied on BERT (which turned out to be still competitive, even against the LLM-based baselines), to keep the approach lightweight (Gao et al., 2025) and to avoid the self-preference bias, where LLMs tend to score their own outputs higher than others (Panickssery et al., 2024).

Ethical Considerations

Ethical approval for the human experiments conducted in this study was obtained from the Ethics Board at Utrecht University. The 100 crowdworkers recruited on Prolific were paid £1 for an estimated workload of 10 minutes, which corresponds to £6 per hour, matching the minimum pay according to Prolific. They gave informed consent before participating in the experiment.

Supplementary Materials Availability Statement: The code to reproduce the results in this paper and the formulaicness dataset are available on GitHub at: `https://github.com/Eduardo-Calo/formulaicness`. The BERT-based regressor is available on Hugging Face at: `https://huggingface.co/Eduardo-Calo/formulaicness`.

Acknowledgments

We thank Joël Weber for contributions to an earlier version of this work.

References

Dave Barker-Plummer, Richard Cox, and Robert Dale. 2011. Student translations of natural language into logic: the Grade Grinder Corpus release 1.0. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 51–60.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Eduardo Calò, Elze van der Werf, Albert Gatt, and Kees van Deemter. 2022. Enhancing and evaluating the grammatical framework approach to logic-to-text generation. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 148–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. *arXiv preprint*. ArXiv:2502.15652 [cs].

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233. Place: US Publisher: American Psychological Association.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Isabel Groves, Ye Tian, and Ioannis Douratsos. 2018. Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 109–118, Tilburg University, The Netherlands. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Levon Haroutunian, Zhuang Li, Lucian Galescu, Philip Cohen, Raj Tumuluri, and Gholamreza Haffari. 2023. Reranking for natural language generation from logical forms: A study based on large language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1067–1082, Nusa Dua, Bali. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™ : A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference*

on Natural Language Generation, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free Evaluation Metrics for Text Generation: A Survey. *arXiv preprint*. ArXiv:2501.12011 [cs].

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Rushang Karia, Daniel Bramblett, Daksh Dobhal, Pulkit Verma, and Siddharth Srivastava. 2024. ∀uto∃val: Autonomous Assessment of LLMs in Formal Synthesis and Interpretation Tasks. *arXiv preprint*. ArXiv:2403.18327 [cs].

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA. Section: 12.

Vladimir Iosifovich Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Naturalness evaluation of natural language generation in task-oriented dialogues using BERT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 839–845, Held Online. INCOMA Ltd.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Aikaterini Mpagouli and Ioannis Hatzilygeroudis. 2009. A Knowledge-based System for Translating FOL Formulas into NL Sentences. In Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, editors, *Artificial Intelligence Applications and Innovations III*, volume 296, pages 157–163. Springer US, Boston, MA. Series Title: IFIP Advances in Information and Communication Technology.

Bang Nguyen, Mengxia Yu, Yun Huang, and Meng Jiang. 2024. Reference-based metrics disprove themselves in question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13651–13666, Miami, Florida, USA. Association for Computational Linguistics.

Anna Nikiforovskaya and Claire Gardent. 2024. Evaluating RDF-to-text generation models for English and Russian on out of domain data. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 134–144, Tokyo, Japan. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Barbara Hall Partee, Alice G. B. ter Meulen, and Robert Eugene Wall. 1990. *Mathematical methods in linguistics*. Number v. 30 in Studies in linguistics and philosophy. Kluwer Academic, Dordrecht ; Boston.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarne Ranta. 2011a. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Aarne Ranta. 2011b. Translating between Language and Logic: What Is Easy and What Is Difficult. In Nikolaj Bjørner and Viorica Sofronie-Stokkermans, editors, *Automated Deduction – CADE-23*, volume 6803, pages 5–25. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Alan Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, Hai Wang, and Chris Wroe. 2004. OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. In *Engineering Knowledge in the Age of the Semantic Web*, pages 63–81, Berlin, Heidelberg. Springer.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2024. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Juen-tin Wang. 1980. On computational sentence generation from logical form. In *COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xin Wu, Yi Cai, Zetao Lian, Ho-fung Leung, and Tao Wang. 2023. Generating Natural Language From Logic Expressions With Structural Representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1499–1510.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Erion Çano and Ondřej Bojar. 2020. Automating Text Naturalness Evaluation of NLG Systems. *arXiv preprint*. ArXiv:2006.13268 [cs].

A Details on the Generators

Ranta (Ranta, 2011b) proposes several syntactic transformations to improve fluency in logic-to-text generation in his Grammatical Framework (Ranta, 2011a) implementation. *Flattening* turns nested conjunctions into lists (e.g., *P and Q and R* := *P, Q, and R*). *Aggregation* merges clauses with shared elements (e.g., *x is even or x is odd* := *x is even or odd*). *In-situ quantification* replaces bound variables with quantified phrases (e.g., *for all x: x is even or odd* := *every x is even or odd*). *Verb negation* internalizes negation (e.g., *it is not the case that x is even* := *x is not even*). *Reflexivization* simplifies repeated arguments (e.g., *x equals x* := *x equals itself*). Finally, *modification* combines types and predicates (e.g., *x is a number and x is even* := *x is an even number*).

LoLa (Calò et al., 2022) applies logical equivalence transformations to the input formula before verbalization. These transformations include a range of equivalence laws (Partee et al., 1990) from propositional and first-order logic, such as associativity, commutativity, distributivity, De Morgan's laws, double negation, and vacuous quantification. To apply these optimizations, LoLa constructs a search tree, where each node represents a logically equivalent reformulation of the input. Once the tree is built, all formulas are passed through syntactic optimization and linearized, and the shortest resulting verbalization is returned.

B Details on the Dataset

To build our dataset, we considered only the textual portion of Calò et al. (2022), selecting a balanced subset in which the texts produced by the Literal, Ranta, and LoLa, alongside the original human-written sentences, are equally represented, to account for varying degrees of distance with respect to the input formula. We filtered out texts in which the character-level Levenshtein distance (Levenshtein, 1966) between any two texts derived from the same formula was less than 10, to avoid overly similar texts. We also removed duplicate entries. Table 7 gives some summary statistics on formulaicness bin, and Table 8 some dataset-level statistics. The dataset we ended up with exhibits a total number of 71 unique content words. Refer to Figure 1 for the 20 most frequent ones.

Source	Bin	Avg. Length (Tokens)	Avg. Literals/Text
Literal	[0.0, 0.295)	29.80	6.08
Ranta	[0.103, 0.458)	25.15	5.51
LoLa	[0.424, 0.815)	18.32	4.19
Human	[0.8, 1.0)	11.61	1.58

Table 7: Statistics by original source and bin. Bins were derived from the original TrueSkill's μ and θ values per system (apart from those of Human, which were arbitrarily assigned). Avg. Literals/Text: the average of single-letter characters (i.e., constants and variables) per text.

Split	Size	Avg. Length (Tokens)	Avg. Formulaicness
Train	399	22.19	0.508
Validation	85	21.36	0.513
Test	86	21.92	0.512
Total	570	22.03	0.510

Table 8: Dataset-level statistics.

C Details on Model Training

To fine-tune the BERT-based regressor, we started from the `bert-base-uncased` checkpoint available on Hugging Face. We used the training set to fit the model parameters, the validation set to tune hyperparameters and prevent overfitting (via early stopping), and kept the test set aside for final evaluation. Table 9 shows the set of hyperparameters. All experiments were run on a Tesla T4 GPU (16GB VRAM) with CUDA 12.4 and driver version 550.54.15. See Figure 2 for the training curves.

Hyperparameter	Value
Loss function	Mean Squared Error (MSE)
Optimizer	AdamW
Learning rate	5×10^{-5}
Batch size	8
Epochs	10 (with early stopping)
Dropout	0.1

Table 9: Hyperparameters used for fine-tuning the BERT regressor.

D Details on Model Testing

Since formulaicness is coincidentally correlated with sentence length, i.e., longer sentences tend to be more formulaic (as shown in Table 7), we tested whether the BERT regressor might be relying spuriously on length alone. To do this, we compared the model's predictions against two baselines: a length-based predictor and the ideal identity line (see Figure 3). While the length-based baseline performs reasonably well, the BERT regressor yields

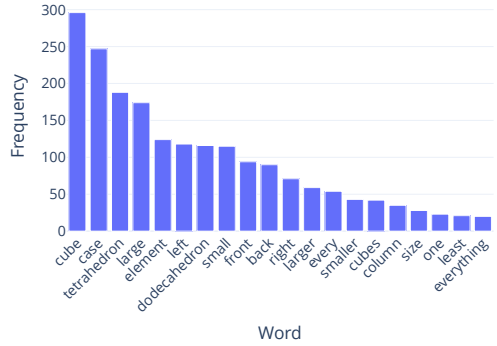


Figure 1: Top 20 most frequent content words in the dataset.

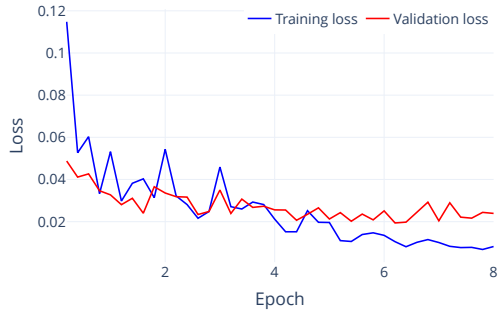


Figure 2: Training and validation loss from BERT regressor fine-tuning.

predictions that more closely follow the ideal identity line, suggesting that the regressor captures additional linguistic nuances beyond sentence length.

To examine which tokens most strongly influence the prediction of the formulaicness regressor, we performed a SHAP values analysis (Lundberg and Lee, 2017). Figure 4 displays the top 20 tokens with the highest average SHAP values. Notably, tokens such as *case*, f, x, and e contribute substantially to the model's predictions of formulaicness. This aligns with expectations: *case* frequently appears in the phrase *it is not the case that*, while letters like f, x, and e are commonly used as logical variables and constants. Given this, we further tested whether the token *case* is blindly assigned high formulaicness (Table 10): when the model is fed texts where *case* appears in non-formulaic contexts,[3] the predicted formulaicness scores are indeed low, suggesting that the model does not rely solely on token presence but also considers con-

[3]Examples from https://en.wiktionary.org/wiki/case.

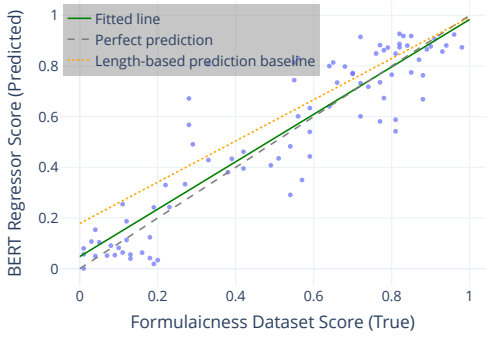


Figure 3: BERT regressor performance against a length-based baseline and the ideal identity line.

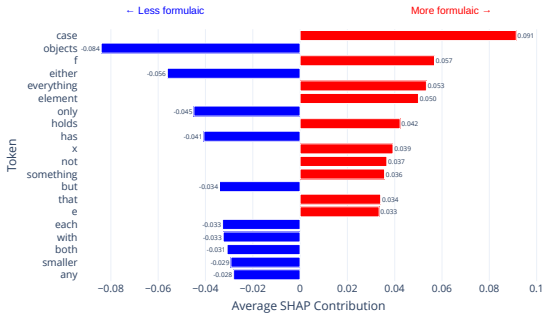


Figure 4: Top 20 tokens with the highest average SHAP values.

text. Moreover, these examples hint that the model can generalize beyond the domain of geometrical shapes (the most prevalent domain in our dataset; Figure 1) likely thanks to the BERT backbone.

E Details on Baseline Metrics

GRUEN GRUEN (Zhu and Bhat, 2020) is a reference-less metric that evaluates text along four dimensions, including grammaticality. Grammaticality is assessed by combining sentence likelihood and grammatical acceptability, two properties that are likely to influence perceptions of naturalness. Sentence likelihood is computed with a BERT-base model, while grammatical acceptability is computed with a fine-tuned BERT model on the Corpus of Linguistic Acceptability dataset (CoLA; Warstadt et al. 2019).

Flesch Reading Ease (FRE) The Flesch Reading Ease score (Flesch, 1948) is a traditional readability metric designed to assess the ease of reading of a text, with higher scores indicating easier-to-read texts. It is based on sentence length and syllable count per word. Low scores may reflect unnat-

Text	Meaning of *case*	Formulaicness
In case of fire, break glass.	An actual situation.	0.17
The doctor told us of an interesting case he had treated that morning.	An instance in a profession.	0.29
The teaching consists of theory lessons and case studies.	An instance as a topic of study.	0.28
The accusative case most commonly indicates a direct object.	Grammatical category.	0.30

Table 10: Sentences containing the word *case* used in non-formulaic contexts, along with the model's predicted formulaicness scores.

ural constructions, such as overly long or clause-heavy sentences, whereas high scores may correspond to more natural language.

Perplexity (PPL) Perplexity (Jelinek et al., 1977) is a standard metric used to evaluate language models, defined as the exponentiated average negative log-likelihood of a sequence under a given language model. Lower perplexity indicates that the model assigns higher probability to the observed text, suggesting greater predictability, an attribute plausibly associated with naturalness.

SLOR SLOR (Syntactic Log-Odds Ratio; Pauls and Klein 2012; Kann et al. 2018) is a metric based on the negative log-likelihood of a sentence. As with PPL, lower SLOR values suggest greater predictability, plausibly associated with naturalness. Specifically, the SLOR score for a sentence is computed as the log probability of the sentence under a given language model, normalized by the unigram log probability and the sentence length. The intuition behind these normalizations is to prevent rare tokens from disproportionately lowering the sentence score and to ensure that shorter sentences are not unfairly favored over equally fluent longer ones.

Llama 3 (LLAMA) LLMs have increasingly been used as automated judges for evaluation tasks (e.g., Zheng et al., 2023; Tan et al., 2024; Bavaresco et al., 2025), including for assessing text quality aspects such as naturalness. We followed this growing trend by adopting an LLM-as-a-judge approach using Llama 3.1 8B Instruct (Grattafiori et al., 2024). We used the quantized version available on Hugging Face: `Meta-Llama-3.1-8B-Instruct-Q8_0.gguf`. We asked the model to rate the naturalness of each text in the held-out test set on a scale from 0 to 100, similarly to the human study (§2). The prompt we used is shown in Figure 5.

F Details on the Results

For computing PPL and SLOR, we used `TinyLlama-1.1B-Chat-v1.0` from Hugging Face. For computing unigram probabilities for SLOR, we used the `wikitext` corpus from Hugging Face.

We used empirical values of α and β, obtained by fitting linear regression models to predict human naturalness from the baseline metric and formulaicness scores. For comparability across metrics, all regression models were trained with `intercept=False`. See Table 11 for the regression coefficients we used as the values of tuned α and β for each metric.

See Figure 6 for the scatterplots with the correlations between metric scores and human naturalness ratings, before and after being combined with formulaicness.

Metric	α	β
GRUEN	0.644	0.356
FRE	0.594	0.406
iPPL	0.484	0.516
iSLOR	0.700	0.300
LLAMA	0.656	0.344

Table 11: Weights α and β learned via linear regression for each baseline metric.

You are participating in an experiment that evaluates the naturalness of English sentences.

Definition of Naturalness:
Naturalness tells us the degree to which a sentence is likely to be produced by a native speaker in the given context or situation.

Instructions:
– Rate each sentence on a scale from 0 (very unnatural) to 100 (very natural).
– Focus only on the sentence structure.
– Do not consider the truth or meaning of the sentence.
– Ask yourself: "Could this sentence have been written by a native speaker?"

Examples:
Sentence: For all x, for all y, if x is a dodecahedron and y is a cube and it is not the case that x is larger than y, then x is of the same size as y.
Rating: 10

Sentence: D is a cube, c is a cube, d is not small and c is not small.
Rating: 55

Sentence: Some dodecahedron is neither large nor small.
Rating: 95

Sentence: {sentence}
Rating:

Figure 5: Prompt shown to the LLAMA baseline used as LLM-as-a-judge.

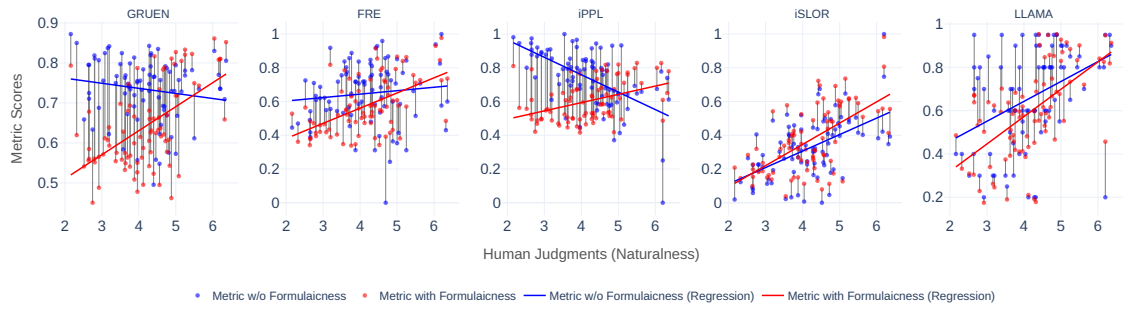


Figure 6: Scatterplots with the correlations between metric scores and human naturalness ratings, before and after being combined with formulaicness.

G Instructions to Participants

Dear participant,

Thank you so much for taking part in this experiment! It will take you approximately 10 minutes to fill in this survey.

If you do wish to participate, your response will be handled anonymously. Collected data will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. If at any point you would like to stop, you can close this form and your response will be deleted.

I have read the above information and understand the purpose of the research and that data will be collected from me. I agree that data gathered for the study may be published or made available, provided my name or other identifying information is not used.

◯ I confirm this.
◯ I do not confirm this and I want to withdraw from participation.

The purpose of this experiment is to assess the quality of some English sentences. We evaluate sentences on the criterion of **naturalness**.

> ***Naturalness*** *tells us the degree to which a sentence is likely to be produced by a native speaker in the given context/situation.*

Please, note down the definition of naturalness, in case you want to refer to it later.

We will present to you around 20 sentences. You will be asked to rate each sentence on a scale from 1 to 7, with 1 being very unnatural, and 7 being very natural.

When rating, you should ask yourself: "Could this sentence have been written by a native speaker?"

While answering the questions, it is important to keep in mind that **we are NOT interested in the meaning of the sentences**. You should **base your opinions on the structure of the sentence only**.

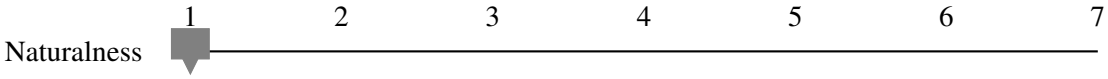
The sentences make claims about a world containing some elements (e.g., a, b, c), as well as about their properties (e.g., cube, small) and relationships (e.g., is in front of).

In the following pages, you will see a couple of guided examples.

Sentence:
For all x, for all y, if x is a dodecahedron and y is a cube and it is not the case that x is larger than y, then x is of the same size as y.

How would you rate the naturalness of the above sentence?

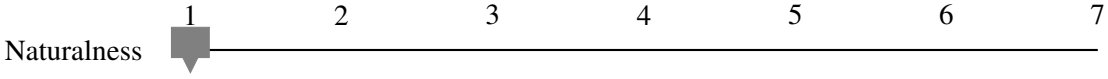
Tip: You might decide that sentences like the one above rate towards the lower end of the sliding bar, because their structure is cumbersome and redundant.



Sentence:
D is a cube, c is a cube, d is not small and c is not small.

How would you rate the naturalness of the above sentence?

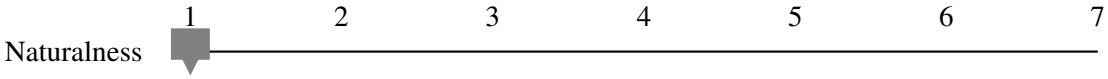
Tip: You might decide that sentences like the one above rate toward the middle of the sliding bar, because their structure is neither overly cumbersome nor particularly straightforward.



Sentence:
Some dodecahedron is neither large nor small.

How would you rate the naturalness of the above sentence?

Tip: You might decide that sentences like the one above rate towards the higher end of the sliding bar, because their structure is clear and straightforward.



Now it is your turn, good luck!