Enhancing and Evaluating the Grammatical Framework Approach to Logic-to-Text Generation

Eduardo Calò[*] and **Elze van der Werf**[*] and **Albert Gatt** and **Kees van Deemter**
Department of Information and Computing Sciences
Utrecht University
Utrecht, the Netherlands
{e.calo, a.gatt, c.j.vandeemter}@uu.nl elzevanderwerf@gmail.com

Abstract

Logic-to-text generation is an important yet underrepresented area of natural language generation (NLG). In particular, most previous works on this topic lack sound evaluation. We address this limitation by building and evaluating a system that generates high-quality English text given a first-order logic (FOL) formula as input. We start by analyzing the performance of Ranta (2011)'s system. Based on this analysis, we develop an extended version of the system, which we name LoLa, that performs formula simplification based on logical equivalences and syntactic transformations. We carry out an extensive evaluation of LoLa using standard automatic metrics and human evaluation. We compare the results against a baseline and Ranta (2011)'s system. The results show that LoLa outperforms the other two systems in most aspects.

https://gitlab.nl4xai.eu/eduardo.calo/LoLa

1 Introduction

Logical formalisms play a pivotal role in many areas of science. Hence, grasping the meaning of these formalisms is crucial for many scholars and researchers. However, this task is not straightforward, and sometimes even experienced logicians might have trouble deciphering a complex formula.

Natural language generation (NLG) techniques can be employed to ease this task. However, logic-to-text generation is understudied, compared to text generation from other inputs (Reiter and Dale, 2000; Gatt and Krahmer, 2018). One notable exception (see §2 for some other examples) is Ranta (2011), a rule-based system that translates between first-order logic (FOL) formulae and natural language (NL). While providing a promising starting point for logic-to-text generation, the system is not evaluated. In our work, we first address this gap via a human translation quality assessment (TQA). Based on this, we propose LoLa, a novel logic-to-text system extending Ranta (2011)'s architecture, which searches for the most suitable formula for translation among the pool of logically equivalent formulae.

We also address one of the many issues that make NLG evaluation challenging (Novikova et al., 2017; Zhou et al., 2022), namely, defining the core dimensions to evaluate (Howcroft et al., 2020), especially issues of meaning vs. grammaticality. These issues come to the fore in logic-to-text generation, where text should be faithful to the original formula, comprehensible, and fluent. These are the central requirements to look for, as text generated from logic can be extremely disfluent and incomprehensible (e.g., a literal translation from a formula), while still being faithful. Furthermore, evaluating faithfulness cannot rely on checking factual accuracy (as in, e.g., WebNLG (Gardent et al., 2017)), due to the problem of *logical form equivalence* (Shieber, 1993), which implies that every formula of FOL is equivalent with infinitely many other FOL formulae, where the question of whether two FOL formulae are logically equivalent is, in general, undecidable. This complicates the problem of finding a formula that is most suitable for being input to an NLG program.

There are also potential trade-offs between evaluation dimensions. For instance, more fluent realizations may sometimes be more ambiguous with respect to a formula, compromising faithfulness (Khan et al., 2012). To use a well-worn example, *Everyone loves someone* can be seen as a correct realization of $\forall x(Person(x) \to \exists y(Person(y) \wedge Love(x, y)))$, but the sentence is ambiguous, also allowing for the more specific interpretation that there exists someone who is loved by everyone (i.e., with the scope of the quantifiers reversed).

[*] These authors contributed equally to this work.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 148 - 171
December 7, 2022 ©2022 Association for Computational Linguistics

In our work, we use a deterministic procedure for generating text from formulae, allowing us to hold faithfulness constant in order to address issues of comprehensibility and fluency.

Our work also addresses the question of which human evaluation task is appropriate for a given system (Gehrmann et al., 2022), proposing a novel evaluation using natural language inference (NLI; Storks et al., 2019; Poliak, 2020).

In outline, the main contributions of this paper are the following:

i) We analyze the quality of the translations of the FOL-to-text system presented in Ranta (2011) via a human translation quality assessment.

ii) We exploit the outcomes of the quality assessment to develop the improved system LoLa.

iii) We present the results of a comprehensive automatic and human evaluation of LoLa.

2 Related Work

Although receiving far less attention than other tasks, generating NL text from (logically rich) meaning representation (MR) formalisms has a relatively long tradition in NLG, with approaches ranging from rule-based (Wang, 1980; Appelt, 1987; Shieber et al., 1990) to statistical (Lu and Ng, 2011; Basile, 2015) and neural models (Wu et al., 2022).

Several MRs have been the focus of the task: logic-based (e.g., description logic (Androutsopoulos et al., 2013), FOL (Mpagouli and Hatzilygeroudis, 2007), and discourse representation structures (Liu et al., 2021; Wang et al., 2021)), graph-based (e.g., Abstract Meaning Representation (AMR; Konstas et al., 2017; Bai et al., 2022, *i.a.*)), and formal languages (e.g., SPARQL (Ngonga Ngomo et al., 2013; Ell et al., 2015)).

Some of these MR formalisms are much simpler than FOL. For example, AMR has less descriptive power (Bos, 2016), whereas datasets such as GeoQuery (Zelle and Mooney, 1996) and RoboCup (Chen and Mooney, 2008), used in, e.g., Wong and Mooney (2007), omit logical operators and variable binding. For these reasons, we select FOL as our formalism, incorporating different types of formulae and defining a concept of *well-behavedness* (see §4.2) to characterize those best suited for logic-to-text translation.

Apart from Ranta (2011), closest to our work are the following approaches. Phillips (1993) considers the problem of logical form equivalence. Mpagouli and Hatzilygeroudis (2009) present a rule-based approach to generate text from FOL with some syntactic optimizations. Coppock and Baxter (2010) propose an algorithm based on dynamic semantics for a specific class of formulae. Kutlak and van Deemter (2015) use background axioms to simplify a FOL formula. Flickinger (2016) generates multiple paraphrases from an input formula. Manome et al. (2018) is one of the few logic-to-text approaches using a sequence-to-sequence framework. Kasenberg et al. (2019) generate explanations from a well-defined logical formalism in the context of human-robot dialogue.

A common thread in most of this work is the absence of (proper) evaluations. In particular, as in Ranta (2011), the proposals in Phillips (1993), Mpagouli and Hatzilygeroudis (2009), Coppock and Baxter (2010), and Kutlak and van Deemter (2015) do not include any attempt at systematic evaluations. In Flickinger (2016), there is a mention of a very preliminary inspection of the paraphrases generated, with pointers for improving the evaluation left for future work. Manome et al. (2018) make an effort to move beyond standard metrics proposing an automatic evaluation based on recognizing textual entailment and present an informal analysis of some generated sentences. However, a proper human evaluation is missing. Kasenberg et al. (2019) do not evaluate their system using automatic metrics. Yet, they perform a human evaluation based on three dimensions and statistically analyze the results. We aim to build a system that, given a logical formula, will produce *effective* texts (i.e., optimally helpful to the needs of the user) (Mayn and van Deemter, 2020), thus, carrying out proper evaluations is one of the main focuses of our work.

3 Model and Data

Ranta (2011) We consider the logic-to-text generation system presented in Ranta (2011)[1] as the starting point for our experiments. The system translates a string from one language into another in two steps: (i) the string in the source language is parsed into an abstract syntax tree (AST), and (ii) the AST is linearized into a string in the target language via language-specific concrete syntax.

The abstract syntax defines functions for several logical constructs, while concrete syntaxes are for-

[1] https://github.com/GrammaticalFramework/gf-contrib/tree/master/cade-2011

mulated to generate FOL linearizations in six NLs. In addition to this, the system performs some *core-to-extended* AST manipulations (e.g., flattening, aggregation, in-situ quantification, verb negation, and reflexivization) to improve fluency. Figure 3 in Appendix A shows a graphical visualization of Ranta (2011)'s system. The system can parse all well-formed FOL formulae without identity (Shapiro and Kouri Kissel, 2021), containing unary and binary predicates and bound variables.

Grade Grinder Corpus The Grade Grinder Corpus (GGC; Barker-Plummer et al., 2011) is a corpus of $> 4.5m$ FOL translations (correct and incorrect) of ca. 300 sentences made by $55k$ students answering exercises in Barwise et al. (2000). Each NL sentence can have multiple (logically equivalent) correct answers.

We select just the portion of answers that are marked as correct and filter the formulae that are not parsable by Ranta (2011)'s system (i.e., formulae with time stamps, mathematical operators, $3-$ and $4-$ary predicates, the identity symbol, and more than 100 characters). This yields around $5,500$ formulae.

Random Generator In GGC, formulae are understandable by humans and have corresponding sentences that are semantically and pragmatically acceptable. However, it might not be representative of the space of all possible formulae. Therefore, we additionally create a tool that generates a random FOL formula in the space of all possible formulae for a given domain lexicon.

4 Assessing Ranta (2011)

To judge the quality of Ranta (2011)'s translation system, we set up a translation quality assessment (TQA; Castilho et al., 2018; Han et al., 2021). A group of human evaluators was asked to analyze a list of English translations from FOL formulae, generated by Ranta (2011)'s system. Specifically, we were interested in receiving feedback on three dimensions: (i) **faithfulness** (i.e., whether the generated text conveys all and only the information of the input formula), (ii) **comprehensibility** (i.e., whether the generated text is clearly understandable by the evaluator), and (iii) **fluency** (i.e., whether the generated text is grammatically accurate and natural-sounding). The evaluation dimensions, especially faithfulness and comprehensibility, are not entirely independent of each other. Nonetheless, they possess their own traits that we wanted to assess separately. A problematic point is establishing faithfulness to an underlying formula in presence of an ambiguous or incomprehensible sentence. To mitigate this problem, in the NLI task introduced for the human evaluation (see §6.2), we gave participants the opportunity to signal text that is ambiguous or incomprehensible.

A total of 10 participants (master students, 4 males and 6 females, with a median age of 24.0 years, $\text{SD} = 1.2$) with sufficient knowledge of logic and proficiency in English[2] voluntarily participated in the study.

Setup Evaluators were shown batches of 25 formula-translation pairs consisting of (i) a random selection of 10 formulae extracted from the parsable portion of the GGC corpus and their Ranta (2011)'s translations, (ii) 10 randomly generated formulae and their Ranta (2011)'s translations, and (iii) 5 filler formulae with incorrect translations created manually. All participants saw the same 5 filler items. The purpose of the fillers was to verify the participants' knowledge of FOL. Were at least 2 out of 5 filler items not identified as such by a participant, their survey response would be ignored entirely in the analysis. None of the participants was omitted by this criterion. See Appendix G for details on the construction of the fillers and Table 13 for the complete list.

To ensure coverage, each participant was presented with a different batch of experimental items.[3] They were required to judge formula-translation pairs under the three dimensions mentioned above. In particular, for each pair, they had to answer a polar question on the translation's faithfulness with the original formula, and rate on a $5-$point Likert scale the translation's comprehensibility and fluency. Moreover, the evaluators were asked to perform *full post-editing* (Hu and Cadwell, 2016) on the translations. The instructions given, the questions asked, and one example batch of experimental items can be found in Appendix G.

[2] The participants involved were students of a MSc in artificial intelligence taught in English, which requires knowledge of logic (in particular, propositional logic and FOL) and proficiency in English for enrolling. Moreover, at the beginning of the questionnaire, participants were required to rate their own knowledge of FOL on a $5-$point Likert scale.

[3] Presenting different batches to each subject prevented us from computing some quantitative analyses, such as inter-annotator agreement. However, the main goal of the TQA was to perform a qualitative study of the system performance and get a more thorough overview of the quality of the translations.

4.1 TQA Results

Evaluators marked 91% of Ranta (2011)'s translations as faithful to the original formula. Most of the translations marked as unfaithful consisted of filler translations. The few non-filler translations that were marked incorrectly were ambiguous and misunderstood by the participants. Thus, we can safely assume that Ranta (2011)'s system, due to its deterministic nature, is robust enough in correctly parsing the structure of the input formula, producing faithful translations.

The average rating of the translations was 3.99 (SD = 1.10) for comprehensibility and 3.26 (SD = 1.32) for fluency. Interestingly, we found that the average faithfulness, comprehensibility, and fluency of the translations from randomly generated formulae are lower than those of the translations from the GGC formulae, as shown in Table 1. We observed a moderate positive correlation (using Pearson's r coefficient) between the comprehensibility and fluency of the (non-filler) translations ($r(198) = 0.60$, $p < .01$). Furthermore, we found a weak negative correlation between formula complexity (in number of connectives, i.e., a formula with more connectives is more complex) and comprehensibility of the corresponding translations ($r(198) = -0.18$, $p < .01$) and a weak negative correlation between formula complexity and fluency of the corresponding translations ($r(198) = -0.24$, $p < .01$). We also observed weak negative correlations between translation length (in number of words) and comprehensibility ($r(198) = -0.23$, $p < .01$), as well as between translation length and fluency ($r(198) = -0.34$, $p < .01$).

Type	#	Faithfulness	Comprehensibility		Fluency	
			μ	σ	μ	σ
GGC	100	93%	4.10	1.02	3.37	1.34
RG	100	88%	3.87	1.15	3.15	1.29

Table 1: The percentage of translations marked as faithful, and the mean (μ) and standard deviation (σ) of the comprehensibility and fluency of translations on a scale of 1 to 5, reported for corpus formulae (GGC) vs. randomly generated formulae (RG).

Post-Edits Post-edits were suggested for 51% of Ranta (2011)'s translations and are often shorter (in word count) than the original translations. The edits can be roughly divided into three categories: (i) syntactic optimizations similar to the *core-to-extended* AST manipulations introduced in Ranta (2011) (see §3), (ii) conversions based on logical equivalences, and (iii) paraphrases using a variety of linguistic constructions. Table 2, Table 3, and Table 9 respectively show some examples for each of the three categories. See Appendix D for a detailed description of these categories.

4.2 Well-Behavedness

Ranta (2011)'s system accepts as input all well-formed FOL formulae (see §3). However, the TQA results suggest that it might be practical to narrow down the definition of formulae suitable for translation to a more restricted subset. The set of well-formed formulae also includes formulae which the participants in the TQA had difficulties with or provided post-edit suggestions for, i.e., formulae with vacuous quantification (e.g., $(\forall x)Even(2)$ or $(\forall x)(\forall x)Even(x)$), formulae with double negation (e.g., $\neg\neg Even(2)$), formulae with nested implication (e.g., $(Odd(1) \rightarrow (Odd(3)) \rightarrow Odd(5))$, and formulae with 8 or more connectives. Translating literally such formulae could result in incomprehensible and disfluent sentences.

Therefore, we operationalize *well-behavedness* as 'the property that a formula should have to be structurally suitable as input for translation into NL'. Well-behavedness is achieved by applying a number of rules to avoid formulae with certain properties, such as double negation and vacuous quantification. The formal definition is present in Appendix B. Formulae that are not well-behaved will be referred to as *ill-behaved*.

5 LoLA

Based on the post-edits we received in the TQA described in §4, we developed LoLA (system for translating between **Lo**gic and **La**nguage), a new logic-to-text system that keeps Ranta (2011)'s original system as its backbone but improves it by extending its algorithm. In particular, LoLA implements the first two categories of post-edits derived from the results of the TQA, i.e., *core-to-extended* AST-like manipulations and logic-based simplification, leaving out the third one (stylistic paraphrases) for future work.

The first class of improvements extends the list of Ranta (2011)'s *core-to-extended* AST conversions with some additional optimizations. See Appendix C for details on their implementation. The second class of improvements manipulates an input FOL formula through the application of logical equivalence laws, based on Partee et al. (1993),

Optimization	Original Formula	Ranta (2011) Translation	Post-Edit
Moving the negation inward	$\neg(\exists x)FrontOf(x,a)$	*It is not the case that there is an element **x** such that **x** is in front of **a**.*	*There is no element in front of **a**.*
In-situ quantification	$(\exists x)Small(x)$	*There is an element **x** such that **x** is small.*	*Something is small.*
Predicate-sharing aggregation	$Larger(d,b) \wedge Larger(e,b) \wedge FrontOf(b,e) \wedge FrontOf(b,d)$	***d** is larger than **b**, **e** is larger than **b**, **b** is in front of **e** and **b** is in front of **d**.*	*Both **d** and **e** are larger than **b**, and **b** is in front of both **e** and **d**.*
Reflexivization	$\neg SameShape(a,a)$	***a** is not of the same shape as **a**.*	***a** does not have the same shape as itself.*

Table 2: Optimizations similar to Ranta (2011) *core-to-extended* AST manipulations suggested in the post-edits of the TQA, with the original formulae and Ranta (2011) translations.

Equivalence Law	Original Formula	Ranta (2011) Translation	Post-Edit
Double negation	$\neg\neg(Medium(a) \vee FrontOf(a,b))$	*It is not the case that it is not the case that **a** is medium or in front of **b**.*	***a** is medium or in front of **b**.*
Redundant information	$\neg SameCol(e,d) \wedge \neg SameCol(e,c) \wedge \neg SameCol(e,d)$	***e** is not in the same column as **d**, **e** is not in the same column as **c** and **e** is not in the same column as **d**.*	***e** is neither in the same column as **d**, nor in the same column as **c**.*
De Morgan's laws	$\neg(Tet(b) \vee Tet(d))$	*It is not the case that **b** is a tetrahedron or **d** is a tetrahedron.*	*Neither **b** nor **d** is a tetrahedron.*
Simplification of $\neg(\exists x)\phi$ to $(\forall x)\neg\phi$	$\neg(\exists y)SameCol(a,y)$	*It is not the case that there is an element **y** such that **a** is in the same column as **y**.*	*All **y**'s are not in the same column as **a**.*

Table 3: Optimizations based on logical equivalence laws suggested in the post-edits of the TQA, with the original formulae and Ranta (2011) translations.

with two additional laws to deal with vacuous quantification. See Table 8 for the list of laws.

The search for the optimal translation is performed as follows. A tree of possible formula manipulations is constructed, with as root node the input formula's AST, and where each node's children are manipulations of the AST that result in a different AST. This tree has a maximum depth because in many cases there are infinitely many manipulations. The maximum depth was experimentally set to 5. After the construction of the search tree, all ASTs in the tree are optimized with the full list of *core-to-extended* AST conversions and linearized, after which the shortest linearization in the tree is returned. The results of the TQA (see §4.1) show that there is a weak negative correlation ($r(198) = -0.23$) between translation length and its assessed comprehensibility but a somewhat stronger negative correlation ($r(198) = -0.34$) between translation length and its assessed fluency. Therefore, we decided to pick the length of the translation (in number of words) as the selection criterion.[4] Figure 4 in Appendix C shows an example of a search tree of formula manipulation sequences.

6 Evaluation

To assess the quality of FOL to NL translations of LoLA, we set up a thorough comparative evaluation experiment. We compared the translation quality of three different systems: (i) a BASELINE generating near-literal translations of formulae, which is Ranta (2011)'s system without its *core-to-extended* AST optimizations, (ii) Ranta (2011), and (iii) LoLA. We run standard automatic NLG metrics based on $n-$gram overlap and semantic similarity, conduct a human evaluation, and compute correlations between the results. The dimensions on which the translation quality of the systems was evaluated are **comprehensibility** and **fluency**.[5] The evaluation also partly focused on well-behaved vs. ill-behaved formulae (see §4.2), investigating how different types of formulae impact the quality of the translations.

6.1 Automatic Evaluation

For the automatic evaluation, we considered all the formulae included in the parsable portion of the GGC (see §3) with their associated ground truth NL references. Each formula was given as input to the three systems to be translated into English.[6] We then compared the realizations of the three systems with the ground truth references. We used seven automatic metrics, three of which are based on $n-$gram overlap, namely, BLEU (Papineni et al., 2002),[7] METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004), two on ELMo embeddings (Peters et al., 2018), namely, Word Mover's Distance (WMD; Kusner et al., 2015)[8] and Sentence Mover's Similarity (SMS; Clark et al., 2019),[9] and two on BERT (Devlin et al., 2019), namely, BERTScore (Zhang et al., 2020),[10] and SBERT (Reimers and Gurevych, 2019).[11] For BERTScore, METEOR, ROUGE-L, and SacreBLEU, we used the implementations provided by Hugging Face (Wolf

[4] If there are multiple shortest linearizations, the first occurrence encountered in a depth-first traversal is chosen.

[5] In contrast to the TQA (see §4), faithfulness was not considered one of the evaluation dimensions because the results of the TQA show that the translations of Ranta (2011) are always faithful. This also holds for BASELINE (since the extended syntax constructs are inherently equivalent to their core syntax counterparts), and remains true for LoLA (since the formula simplifications are based precisely on the laws of logical equivalence).

[6] See Table 10 in Appendix E for some examples.

[7] We used the SacreBLEU (Post, 2018) implementation for improved reproducibility.

[8] https://github.com/src-d/wmd-relax

[9] https://github.com/eaclark07/sms

[10] We used the model roberta-large_L17_no-idf.

[11] We computed cosine similarity after obtaining sentence embeddings with the model all-distilroberta-v1.

et al., 2020).[12] Table 4 summarizes the results obtained.

System	$n-$gram-based Metrics			Semantics-based Metrics			
	METEOR	ROUGE-L	SacreBLEU	BERTScore	SBERT	SMS	WMD
BASELINE	46.66	31.46	9.69	88.10	72.18	6.30	1.19
Ranta (2011)	50.10	36.54	11.70	89.00	72.93	23.68	14.69
LOLA	**53.87**	**45.01**	**17.27**	**90.77**	**77.89**	**54.11**	**38.92**

Table 4: Performance of the three systems against the GGC ground truth references according to the automatic metrics. All scores are reported on the same scale to improve readability.

LOLA outperforms the other two systems on all metrics. However, in the context of logic-to-text generation, the results of metrics based on $n-$gram overlap vs. metrics based on semantic similarity should be interpreted differently. The texts that we are comparing (i.e., GGC ground truth references and texts generated by the three systems) differ considerably in their structural realization, while keeping the same underlying meaning (i.e., they are paraphrases). This is due to the fact that the GGC ground truth references explain the logical formulae, which, in turn, are given as input to the three systems that operate in deterministic ways, ensuring faithfulness of the output texts. Therefore, we expect the results of metrics based on semantic similarity to be comparable across the three systems. On the contrary, we should notice more variance with metrics based on $n-$gram overlap, since they are more reliant on the surface structure of the texts.

Nevertheless, BERTScore is the only semantics-based metric that is close to following the expected behavior. This might be an additional indication that neural language models are not capable yet to capture deep semantics of NLs, but are still biased towards morphosyntactic realizations (Bender and Koller, 2020). On the other hand, we can observe substantial variance in the results involving $n-$gram-based metrics. All these metrics favor LOLA, which apparently creates texts structurally closer to the original GGC ground truth references.

6.2 Human Evaluation

The human evaluation consisted of two tasks: (i) a natural language inference (NLI) task to assess the comprehensibility of the translations and (ii) a fluency ranking (FR) task (Bojar et al., 2014) to assess the fluency of the translations. The instructions given and the questions asked to the participants can be found in Appendix H.

Half of the formulae used in the experimental items were extracted from the GGC, while the other half were randomly generated. This resulted in a set of formulae that contained both well-behaved and ill-behaved formulae and was representative of the entire space of FOL formulae. Each formula was given as input to the three systems to be translated into English.

A total of 21 participants (researchers and students, 9 males and 12 females, with a median age of 25.0 years, SD $= 12.2$) with sufficient knowledge of logic and proficiency in English[13] were recruited for the task.

Setup The rationale for using NLI is that it taps comprehension, allowing us to gauge the extent to which FOL translations by different systems facilitate inference. In this case, the fact that the underlying meaning is captured by a logical formula ensures that the task is highly controlled. Additionally, NLI allows checking more objectively how well participants understand text, removing the factor of subjectivity that characterizes other evaluation methods such as Likert scales.

The NLI task was framed in such a way that the three system translations (one per system) of the same formula were considered as *premises* associated with the same *hypothesis* (manually crafted) each time. An illustration of this is presented in Table 5. The third answer option *Other* was added for cases in which the premise was ambiguous or unclear for the participant.[14]

Participants were randomly assigned to one of three groups. Items for the experiment (where an item consists of a formula translation by only one of the three systems and the associated hypothesis) and participant groups were counterbalanced by rotating through a 3 (system) × 3 (participant group) Latin square (Fisher, 1925). This ensured that the experimental items were counterbalanced, so that every item was shown to approximately the same number of participants and every participant was shown the same number of items (42), while participants only saw one system translation per

[12] https://huggingface.co/evaluate-metric

[13] All participants had taken at least one course on FOL. Furthermore, at the beginning of the questionnaire, participants were asked to rate their knowledge of logic on a $4-$point Likert scale and their proficiency in English on a $5-$point Likert scale.

[14] In the analysis, the *Other* option was always marked incorrect because ambiguous and unclear translations are less understandable.

formula.

BASELINE	Does the hypothesis automatically follow from the premise? Premise: *"b is a cube or it is not the case that b is a cube and c is a cube."* Hypothesis: *"only c is a cube."* Yes | No | Other ________
Ranta (2011)	Does the hypothesis automatically follow from the premise? Premise: *"All these hold:* *- b is a cube or b is not a cube;* *- c is a cube."* Hypothesis: *"only c is a cube."* Yes | No | Other ________
LoLA	Does the hypothesis automatically follow from the premise? Premise: *"c is a cube."* Hypothesis: *"only c is a cube."* Yes | No | Other ________

Table 5: Example of three NLI experimental items derived from translating the formula $(Cube(b) \vee \neg Cube(b)) \wedge Cube(c)$ with the three systems.

The motivation behind using FR is that evaluating fluency in an absolute manner can be tricky. Comparing different outputs can aid evaluators to make more informed judgments. In this task, participants were asked to rank the translations of the three different systems of the same source formula according to the criterion of fluency. Ties were allowed. An illustration of this is presented in Figure 1. The FR task did not require a Latin square design because all three translations per formula were presented together in the same experimental item. Therefore, each group of participants was shown the same set of 20 FR questions.

Given the following formula and candidate translations, rank the translations from most fluent to least fluent.

Formula: $(Cube(b) \vee \neg Cube(b)) \wedge Cube(c)$
System 1: *"b is a cube or it is not the case that b is a cube and c is a cube."*
System 2: *"All these hold:*
- b is a cube or b is not a cube;
- c is a cube."
System 3: *"c is a cube."*

Figure 1: An illustration of a FR question in the experiment.

NLI Results The comprehensibility of a translation was calculated as the proportion of correct answers (i.e., correctly spotted presence or absence of entailment) to its corresponding NLI question. The mean of the percentage of correct NLI answers per participant was 70.4% (SD $=$ 8.9%) and the mean of the percentage of correct answers per question was 70.2% (SD $=$ 28.7%). The interannotator agreement was very low (Krippendorff's $\alpha = 0.181$), highlighting the difficulty of this task. Two outlier NLI questions, on which the participants performed significantly worse than on other questions, with the percentage of correct answers being more than two standard deviations below the mean, were removed from the analysis.

The translations from LoLA had the highest mean of the percentage of correct answers. Figure 2 shows the distribution of the percentages of correct participant responses for the different types of formulae.[15] A two-way system × well-behavedness ANOVA revealed that there was a significant interaction between the effects of translation system and formula type on the percentage of correct answers ($F(2, 114) = 3.11$, $p = .048$). Simple main effects analysis showed that translations from well-behaved formulae received a significantly higher percentage of correct answers than translations from ill-behaved formulae ($F(1) = 8.87$, $p = .004$), and that there was a significant effect of translation system on the percentage of correct answers ($F(2) = 5.50$, $p = .005$).

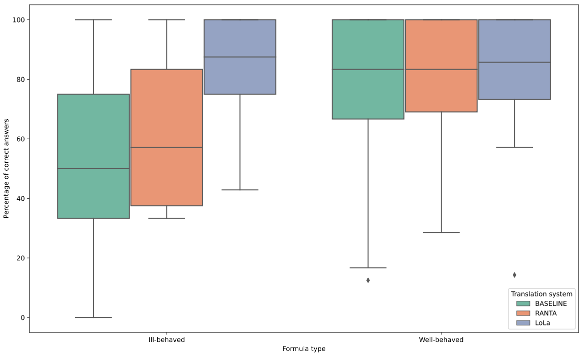


Figure 2: The distribution of the percentage of correct answers to the NLI questions (outliers excluded), grouped by formula type and translation system, as a boxplot showing the medians, lower quartiles, and upper quartiles, along with extreme values. Diamonds are outlier values.

Tukey's HSD test for multiple pairwise comparisons showed that the effect of translation system is mainly due to a difference between BASELINE and LoLA ($p = .005$, $+16.59$ under LoLA). There were no significant differences between BASELINE and Ranta (2011) ($p = .657$) and LoLA and Ranta (2011) ($p = .053$). Tukey's HSD test revealed also that the interaction effect found is mainly due to

[15] Note that this figure shows some outliers other than the ones removed from the analysis. The outliers removed from the analysis were the questions with a mean of the percentage of correct answers more than two standard deviations below the mean over all the questions, while the outliers in this figure are the outlier questions per formula type per translation system.

the extremely low percentage of correct answers for Baseline translations from ill-behaved formulae. All three translation systems had a higher percentage of correct answers for well-behaved formulae: Baseline ($p = .021$, $+23.75$ under Baseline's translations from well-behaved formulae), Ranta (2011) ($p = .013$, $+24.97$ under Ranta (2011)'s translations from well-behaved formulae), and LoLa ($p = .002$, $+29.52$ under LoLa's translations from well-behaved formulae). Furthermore, LoLa's translations from ill-behaved formulae were higher than Baseline's translations from ill-behaved formulae ($p = .002$, $+8.19$ under LoLa).

FR Results To calculate the ranking of the three systems based on the individual rankings the participants gave in each of the FR questions, we used the TrueSkill adaptation of Sakaguchi et al. (2014).[16] TrueSkill was run 200 times on 1260 pairwise rankings derived from the 420 collected system rankings (20 per participant; Krippendorff's $\alpha = 0.475$). The results of the clustering of systems with overlapping rank ranges are presented in Table 6. There were significant differences between the ranked fluency of the three systems, such that they were all in a different cluster. The final ranking was LoLa > Ranta (2011) > Baseline.

#	μ	Rank Range	System
1	3.539	$1-1$	LoLa
2	-0.643	$2-2$	Ranta (2011)
3	-2.873	$3-3$	Baseline

Table 6: The final ranking of the three systems according to TrueSkill (significance cluster number at $p-$level $p \leq .02$ (#), the final estimate of the system's ability (μ; inferred mean), the range of ranks in which the system falls, and system name).

To test whether there was an interaction effect on ranked fluency between the type of formulae (well-behaved or ill-behaved) and translation system, TrueSkill was run for well-behaved and ill-behaved formulae separately. For well-behaved formulae, the model was run 200 times on 693 pairwise collected rankings derived from the 231 system rankings (11 per participant). For ill-behaved formulae, the model was run 200 times on 567 pairwise collected rankings derived from the 189 system rankings (9 per participant). The results of the clustering of systems with overlapping rank ranges for well-behaved formulae vs. ill-behaved formulae are presented in Table 7. We found a difference between the fluency of translations from well-behaved formulae by Ranta (2011) vs. LoLa, in addition to a difference in fluency between the two systems for translations from ill-behaved formulae (in both cases, LoLa had a higher rank than Ranta (2011)).

Well-Behaved Formulae				Ill-Behaved Formulae			
#	μ	Rank Range	System	#	μ	Rank Range	System
1	2.256	$1-1$	LoLa	1	3.826	$1-1$	LoLa
2	0.325	$2-2$	Ranta (2011)	2	-1.423	$2-2$	Ranta (2011)
3	-2.594	$3-3$	Baseline	3	-2.355	$3-3$	Baseline

Table 7: The final rankings of the three systems for well-behaved formulae vs. ill-behaved formulae according to TrueSkill (significance cluster number at $p-$level $p \leq .02$ (#), the final estimate of the system's ability (μ; inferred mean), the range of ranks in which the system falls, and system name).

6.3 Correlations between Automatic Metrics and Human Judgments

In order to have a more comprehensive picture of our experiments, we calculated correlations (using Pearson's r coefficient) between the results of the automatic evaluation and the judgments obtained during the human evaluation. We considered only the experimental items derived from the formulae extracted from the GGC used for the NLI and FR tasks, as they have ground truth NL references and thus were scored using the automatic metrics. As for the metrics, we considered BERTScore, ROUGE-L, and SBERT.

For computing the correlation on the NLI task, we calculated a normalized score on $[0, 1]$ per translation, based on the answers given by the participants, such that the closer the score is to 1, the more comprehensible the translation. Similarly, the higher the score given by the automatic metrics to the translation, the more similar (structurally or semantically) to the ground truth reference it is. We observed no correlation between human judgments and any of the metrics ($r(55) = 0.05$, $p = .737$ with BERTScore; $r(55) = -0.08$, $p = .540$ with ROUGE-L; $r(55) = 0.009$, $p = .947$ with SBERT). For reference, Figure 5 in Appendix E shows the scatterplots.

For computing the correlation on the FR task, we scored each translation based on the average ranking received in the FR task. Each translation was scored from 1 (most fluent) to 3 (least fluent) by the participants, so we obtained translations

[16] www.github.com/keisks/wmt-trueskill. See Appendix F for a high-level description of TrueSkill.

scored on $[1, 3]$. In this case, the higher the average FR score, the more the translation is likely to be disfluent. Conversely, the higher the score of the automatic metrics, the more similar to the ground truth reference the translation is. Allegedly, a more fluent translation for humans should receive a higher score from the automatic metrics, so we expect a negative correlation between the average FR scores and the metric scores. This is supported by the results, where we observe weak to moderate negative correlations ($r(25) = -0.48$, $p = .01$ with BERTScore; $r(25) = -0.51$, $p < .01$ with ROUGE-L; $r(25) = -0.35$, $p = .08$ with SBERT), statistically significant for two out of three metrics (BERTScore and ROUGE-L). Figure 6 in Appendix E shows the scatterplots representing the negative correlations.

We proceeded with a manual analysis to further inspect the misalignments between automatic scores and human judgments in the two evaluation tasks. Specifically, we wanted to study two extreme cases, namely, when a high score from the automatic metrics corresponded to poor human judgments, and vice versa. See Appendix E for a detailed report.

Our results highlight an apparent lack of appropriate metrics to automatically evaluate the task of logic-to-text generation, as the metrics that we considered measure different properties than the dimensions we are interested in assessing. Semantics-based metrics focus exclusively on computing semantic similarity between texts. Moreover, they should theoretically be solid enough to capture nuances in meaning, yet we saw that this is mostly not the case (with BERTScore being the only exception). Consequently, the nature of these metrics does not allow them to tackle the core issue of comprehensibility, i.e., whether a text is more understandable than another. Our results also suggest that these metrics are of limited use for assessing fluency. Similarly, metrics relying only on $n-$gram overlap are unsuitable for any task involving comprehension, as they simply compare surface realizations of texts. On the other hand, they might be slightly more appropriate to evaluate fluency, as overlapping tokens can be an indication of fluency.

7 Future Work

We see two main areas for further investigation. First, we will examine to what extent our approach can be scaled up to include FOL with identity by enhancing the generator, the logical equivalence laws, and crucially, the optimization operations that were applied to the sentences generated. Second, we will try to implement the list of linguistic improvements that emerged from the TQA (see Table 9) by investigating methods to programmatically exploit paraphrasing techniques (rule-based, neural, or hybrid), and adequately scoring the resulting translations.

8 Conclusion

We conducted a human TQA on the faithfulness, comprehensibility, and fluency of Ranta (2011)'s translations. We implemented part of the results to build LoLA, an enhanced version of Ranta (2011)'s FOL-to-text system, which optimizes the input formula when generating text. We evaluated LoLA against a baseline and Ranta (2011)'s original system, performing both automatic and human evaluations. Our results suggest that Ranta (2011)'s framework, once adequately enhanced with logical equivalence laws, lends itself well to generating NL translations of FOL formulae. Furthermore, the results indicate the inappropriateness of current standard automatic metrics to evaluate logic-to-text generation, as they focus on assessing different properties than the dimensions relevant for this task.

The present work on logic-to-text generation can be potentially beneficial for a variety of applications. Paraphrasing systems could profit from the constraints given by logical equivalence laws to generate faithful paraphrases. Logic teaching can benefit by incorporating LoLA in an intelligent tutoring system supporting students and educators. LoLA could also be the base for a system helping engineers comprehend the convoluted outputs of theorem provers, as literally translating those formulae might result in quite cumbersome text.

We hope that this paper will motivate researchers in the broader NLG community to focus more on the issue of generating faithful, comprehensible, and fluent text from logically rich inputs.

Limitations

At present, both Ranta (2011) and LoLA do not cover identity (=). When identity is added to FOL, the expressive power of FOL increases very significantly, allowing it to express things like "there are more/fewer than n A's", "exactly n A's are B's", and so on, often using formulae whose structure is very distant from those of normal English sentences.

The current experimental design of the NLI task does not allow us to get fine-grained insights on ambiguity (i.e., the different readings that a translation may induce), which is crucial to avoid misunderstandings about the original meaning of a formula. In particular, the choice of the *Other* option revealed that the participants did spot the existence of ambiguities in the premises, or did not detect them at all, resulting in different interpretations of the premise.

The vocabulary of entities and relations from the GGC is limited in nature, given its pedagogical origin. Enlarging and diversifying the language domain would raise complications such as dealing with logical properties of the predicates, both in isolation and compositionally, implicatures, and world knowledge. Consequently, ensuring the creation of a fair and proper evaluation, especially for the NLI task, would be significantly more challenging.

Our evaluation focused exclusively on English. However, studying this subject from the perspectives of (typologically) different languages would bring up an incredibly wide range of research questions, e.g., is the concept of well-behavedness language-independent? Do the modifications performed to Ranta (2011)'s system scale up to other languages?

Ethics Statement

In the TQA, all the participants (10 master students) agreed to voluntarily participate. In the human evaluation, we paid 14 of the 21 evaluators (researchers and students) €5 upon completion of the survey. The other 7 agreed to participate without remuneration. In both studies, all the participants gave their informed consent to participate anonymously.

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments and suggestions.

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860621.

References

Ion Androutsopoulos, Gerasimos Lampouras, and Dimitrios Galanis. 2013. Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System. *Journal of Artificial Intelligence Research*, 48:671–715.

Douglas E. Appelt. 1987. Bidirectional grammars and the design of natural language generation systems. In *Theoretical Issues in Natural Language Processing 3*.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Jon Barwise. 1977. An introduction to first-order logic. In *Studies in Logic and the Foundations of Mathematics*, volume 90, pages 5–46. Elsevier.

Jon Barwise, John Etchemendy, Gerard Allwein, Dave Barker-Plummer, and Albert Liu. 2000. *Language, Proof and Logic*. CSLI publications.

Valerio Basile. 2015. *From logic to language: Natural language generation from logical forms*. Ph.D. thesis, University of Groningen.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Johan Bos. 2016. Squib: Expressive power of Abstract Meaning Representations. *Computational Linguistics*, 42(3):527–535.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment*, pages 9–38. Springer.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 128–135, New York, NY, USA. Association for Computing Machinery.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Elizabeth Coppock and David Baxter. 2010. A translation from logic to english with dynamic semantics. In *New Frontiers in Artificial Intelligence*, pages 197–216, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Basil Ell, Denny Vrandečić, and Elena Simperl. 2015. SPARTIQULATION: Verbalizing SPARQL Queries. In *The Semantic Web: ESWC 2012 Satellite Events*, Lecture Notes in Computer Science, pages 117–131, Berlin, Heidelberg. Springer.

Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub.

Ronald Aylmer Fisher. 1925. *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Loyd.

Dan Flickinger. 2016. Generating English paraphrases from logic. *From Semantics to Dialectometry*.

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.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. 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.

Ke Hu and Patrick Cadwell. 2016. A comparative study of post-editing guidelines. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 346–353.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, Meia Chita-Tegmark, and Matthias Scheutz. 2019. Generating justifications for norm-related agent decisions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 484–493, Tokyo, Japan. Association for Computational Linguistics.

Imtiaz H. Khan, Kees van Deemter, and Graeme Ritchie. 2012. Managing ambiguity in reference generation: The role of surface structure. *Topics in Cognitive Science*, 4(2):211–231.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Roman Kutlak and Kees van Deemter. 2015. Generating Succinct English Text from FOL Formulae. In *Procs. of First Scottish Workshop on Data-to-Text Generation*, page 3.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2021. Text generation from discourse representation structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 397–415, Online. Association for Computational Linguistics.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Kana Manome, Masashi Yoshikawa, Hitomi Yanaka, Pascual Martínez-Gómez, Koji Mineshima, and Daisuke Bekki. 2018. Neural sentence generation from formal semantics. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 408–414, Tilburg University, The Netherlands. Association for Computational Linguistics.

Alexandra Mayn and Kees van Deemter. 2020. Towards generating effective explanations of logical formulas: Challenges and strategies. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 39–43, Dublin, Ireland. Association for Computational Linguistics.

Aikaterini Mpagouli and Ioannis Hatzilygeroudis. 2007. Converting first order logic into natural language: A first level approach. In *Current Trends in Informatics: 11th Panhellenic Conference on Informatics, PCI*, pages 517–526.

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.

Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. 2013. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 977–988, Rio de Janeiro, Brazil. ACM Press.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Barbara BH Partee, Alice G ter Meulen, and Robert Wall. 1993. *Mathematical methods in linguistics*, volume 30 of *Studies in Linguistics and Philosophy (SLAP)*. Springer Dordrecht.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

John D. Phillips. 1993. Generation of text from logical formulae. *Machine Translation*, 8(4):209–235.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Aarne Ranta. 2011. Translating between language and logic: what is easy and what is difficult. In *Proceedings of the International Conference on Automated Deduction*, pages 5–25. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

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.

Stewart Shapiro and Teresa Kouri Kissel. 2021. Classical Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Stuart M. Shieber. 1993. The problem of logical form equivalence. *Computational Linguistics*, 19(1):179–190.

Stuart M. Shieber, Gertjan van Noord, Fernando C. N. Pereira, and Robert C. Moore. 1990. Semantic-head-driven generation. *Computational Linguistics*, 16(1):30–42.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.

Chunliu Wang, Rik van Noord, Arianna Bisazza, and Johan Bos. 2021. Evaluating text generation from discourse representation structures. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 73–83, Online. Association for Computational Linguistics.

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

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Yuk Wah Wong and Raymond Mooney. 2007. Generation by inverting a semantic parser that uses statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 172–179, Rochester, New York. Association for Computational Linguistics.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1050–1055. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

A Details on Ranta (2011)

Figure 3 shows a graphical schematization of Ranta (2011)'s translation system.

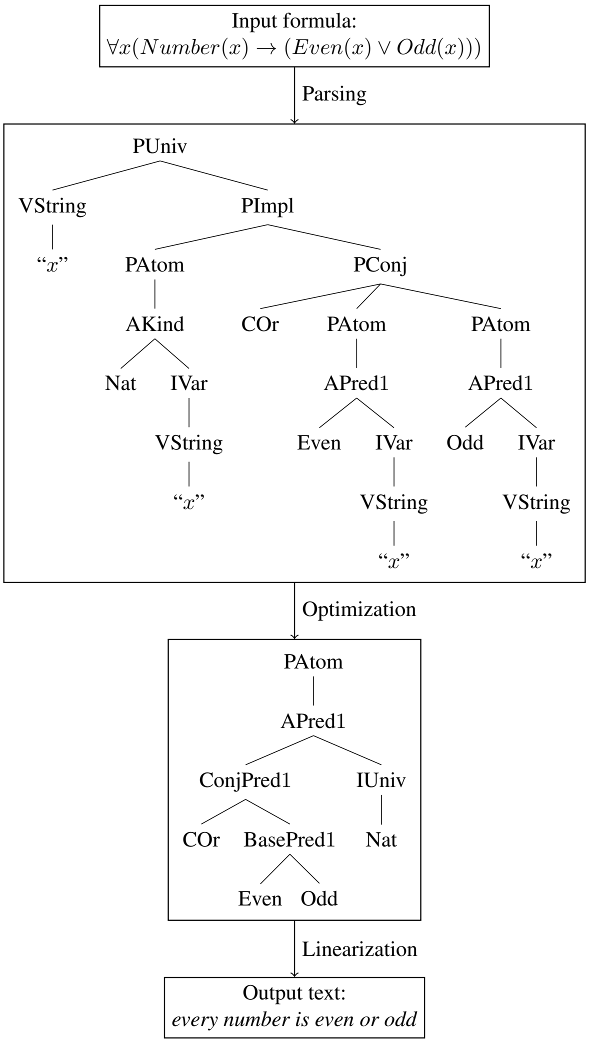


Figure 3: A model of Ranta (2011)'s translation system, with an example translation of a FOL formula into English. Each AST node is named after the syntactic function used to construct the constituent.

B Formal Definition of Well-Behavedness

The following is the formal definition of *well-behavedness*, stating all the conditions that a formula should have to be suitable for translation into NL:

1. All atomic propositions are well-behaved formulae.
2. Negation: if ϕ is a well-behaved formula and it does not contain subformulae of the form $\neg\psi$ for any formula ψ, then $\neg\phi$ is a well-behaved formula.
3. Conjunction: if ϕ and ψ are well-behaved formulae, then $(\phi \wedge \psi)$ is a well-behaved formula.
4. Disjunction: if ϕ and ψ are well-behaved formulae, then $(\phi \vee \psi)$ is a well-behaved formula.
5. Implication: if ϕ and ψ are well-behaved formulae and neither of them has any subformulae of the form $\alpha \rightarrow \beta$ for any set of formulae $\{\alpha, \beta\}$, then $(\phi \rightarrow \psi)$ is a well-behaved formula.
6. Universal quantification: if ϕ is a well-behaved formula, x is a variable, and ϕ contains at least one free occurrence of x, then $(\forall x)\phi$ is a well-behaved formula.
7. Existential quantification: if ϕ is a well-behaved formula, x is a variable, and ϕ contains at least one free occurrence of x, then $(\exists x)\phi$ is a well-behaved formula.
8. Bounded quantification: if ϕ is a proposition, x is a variable, K is a kind predicate, and ϕ contains at least one free occurrence of x, then $(\forall x : K)\phi$ and $(\exists x : K)\phi$ are well-behaved formulae.
9. Conjunction and disjunction of proposition lists: if $\phi_1, ..., \phi_n$ are propositions, then $\wedge[\phi_1, ..., \phi_n]$ and $\vee[\phi_1, ..., \phi_n]$ are propositions.
10. Nothing else is a well-behaved formula.

In addition to this definition, the well-behavedness of a formula also depends on its complexity, calculated in the number of connectives. Only if a formula contains < 8 connectives, it is considered well-behaved.

C Additional Details on LoLa

The list of *core-to-extended* AST conversions was expanded with the following optimization rules.

The rule of *existential negation* turns a negated existential quantifier into a *negative existential*, which asserts the non-existence of an element in the domain of quantification. This optimization should improve translations such as *it is not the case that there exists an element x such that [...]* to *there exists no element x such that [...]*, pushing the negation inward.

The rule of *in-situ quantification without a kind predicate* applies to the special case of *kind predicates* such as *natural number* that, according to Ranta (2011), serve to restrict the domain of quantification. This rule replaces an occurrence of a bound variable in the quantified proposition (∀, ∃, or ⊥) with simpler expressions (*everything*, *some-*

Propositional Logic	First-Order Logic
Idempotence	Quantifier Negation
$P \vee P \Leftrightarrow P$	$\neg(\forall x)\phi(x) \Leftrightarrow (\exists x)\neg\phi(x)$
$P \wedge P \Leftrightarrow P$	$(\forall x)\phi(x) \Leftrightarrow \neg(\exists x)\neg\phi(x)$
	$\neg(\forall x)\neg\phi(x) \Leftrightarrow (\exists x)\phi(x)$
Associativity	$(\forall x)\neg\phi(x) \Leftrightarrow \neg(\exists x)\phi(x)$
$(P \vee Q) \vee R \Leftrightarrow P \vee (Q \vee R)$	
$(P \wedge Q) \wedge R \Leftrightarrow P \wedge (Q \wedge R)$	Quantifier Distribution
	$(\forall x)(\phi(x) \wedge \psi(x)) \Leftrightarrow (\forall x)\phi(x) \wedge (\forall x)\psi(x)$
Commutativity	$(\exists x)(\phi(x) \vee \psi(x)) \Leftrightarrow (\exists x)\phi(x) \vee (\exists x)\psi(x)$
$P \vee Q \Leftrightarrow Q \vee P$	
$P \wedge Q \Leftrightarrow Q \wedge P$	Quantifier Independence
	$(\forall x)(\forall y)\phi(x, y) \Leftrightarrow (\forall y)(\forall x)\phi(x, y)$
Distributivity	$(\exists x)(\exists y)\phi(x, y) \Leftrightarrow (\exists y)(\exists x)\phi(x, y)$
$(P \vee Q) \wedge (P \vee R) \Leftrightarrow P \vee (Q \wedge R)$	
$(P \wedge Q) \vee (P \wedge R) \Leftrightarrow P \wedge (Q \vee R)$	Quantifier Movement
	$\phi \to (\forall x)\psi(x) \Leftrightarrow (\forall x)(\phi \to \psi(x))$
Identity	(if x is not free in ϕ)
$P \vee \bot \Leftrightarrow P$	$\phi \to (\exists x)\psi(x) \Leftrightarrow (\exists x)(\phi \to \psi(x))$
$P \vee \top \Leftrightarrow \top$	(if x is not free in ϕ)
$P \wedge \bot \Leftrightarrow \bot$	$(\forall x)\psi(x) \to \phi \Leftrightarrow (\exists x)(\psi(x) \to \phi)$
$P \wedge \top \Leftrightarrow P$	(if x is not free in ϕ)
	$(\exists x)\psi(x) \to \phi \Leftrightarrow (\forall x)(\psi(x) \to \phi)$
Complement	(if x is not free in ϕ)
$P \vee \neg P \Leftrightarrow \top$	
$\neg\neg P \Leftrightarrow P$ (double negation)	Vacuous Quantification
$P \wedge \neg P \Leftrightarrow \bot$	$(\forall x)\phi \Leftrightarrow \phi$ (if x is not free in ϕ)
	$(\exists x)\phi \Leftrightarrow \phi$ (if x is not free in ϕ)
De Morgan	
$\neg(P \vee Q) \Leftrightarrow \neg P \wedge \neg Q$	
$\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q$	
Conditional	
$P \to Q \Leftrightarrow \neg P \vee Q$	
$P \to Q \Leftrightarrow \neg Q \to \neg P$ (contraposition)	

Table 8: List of logical equivalence laws used as formula conversions in LoLa, where P, Q, or R stand for any arbitrarily chosen well-formed formula, and $\phi(x)$ or $\psi(x)$ for any formula in which x is free.

thing, or *nothing*). As an example, *for all x, x is even* would be optimized to *everything is even*.

The rules for $2-$*place predicate-sharing aggregation* are of two kinds: in *subject-sharing*, different occurrences of the same predicate in a formula share the first argument; in *object-sharing*, different occurrences of the same predicate share the second argument. In these cases, the formula is flattened to merge the occurrences of the predicate. For example, $Parallel(a, b) \wedge Parallel(c, b)$ would be translated as *a and c are parallel to b*.

Finally, the optimization rule to perform *reflexivization* on negated predicates improves translations such as *x is not bigger than x* to *x is not bigger than itself*.

Given the optimization rules presented in this section and in §3 and the equivalence laws in Table 8, the selection of the optimal translation is performed as described in §5. Figure 4 shows an example of a search tree.

D Details on the TQA's Post-Edits

The post-edits suggested by the participants in the TQA (see §4.1) can be divided into three categories. The first category is in the spirit of Ranta (2011)'s *core-to-extended* AST manipulations. The optimizations that the evaluators suggested are: moving the negation inward if an existential quantifier is negated, in-situ quantification without a kind predicate present, two-place predicate-sharing aggregation, and reflexivization of negated predicates. Table 2 shows one example for each suggested optimization.

The second category includes conversions based on the structural manipulation of the form of the input using logical equivalence laws. This way, logically equivalent but arguably more comprehensible and fluent translations can be obtained. Examples of this are the elimination of double negation, the use of De Morgan's laws (Barwise, 1977), and the simplification of $\neg(\exists x)\phi$ to $(\forall x)\neg\phi$. Table 3 presents some examples of conversions suggested in this category.

The third category consists of linguistic and stylistic optimizations of the translations, introducing a greater variety of terms, expressions, and syntactic constructions than those employed by Ranta (2011). Examples of some linguistic constructions introduced are relative clauses, anaphoric expressions, periphrastic expressions, and the rephrasing of connectives. Table 9 presents the complete list of linguistic constructions, together with the logical constructs they can convey.

E Details on the Evaluation

Table 10 presents some outputs generated by the three systems we compared for evaluation. The table highlights the different operations to translate formulae into text used by the three systems. Note, in particular, the convoluted nature of the quasi-literal translations of BASELINE, and the techniques employed by Ranta (2011) and LOLA to improve them. Specifically, Ranta (2011) implements some common techniques in NLG (e.g., aggregation in (3)), while LOLA additionally employs logical equivalence laws (e.g., double negation in (1)) to further refine the translations.

Figure 5 presents the scatterplots showing the relationships (not statistically significant) between the average NLI score and the scores assigned by the automatic metrics to the translations (BERTScore in Figure 5a, ROUGE-L in Figure 5b, and SBERT in Figure 5c). Figure 6 presents the scatterplots showing the weak to moderate negative correlations (statistically significant for BERTScore and ROUGE-L) between the average FR ranking and the scores assigned by the automatic metrics to the translations (BERTScore in Figure 6a, ROUGE-L in Figure 6b, and SBERT in Figure 6c). Note, however, that the FR rank alone (1, 2, or 3) of a translation might not be ideal to measure its fluency. Given that ties were allowed in the FR task, it might be the case that all the translations of an input formula receive a 1. However, this might mean that the translations are all equally disfluent. Therefore, two translations of different formulae receiving a 1 cannot be viewed as equally fluent. Nonetheless, the correlations we found might be due to the fact that few FR rankings resulted in ties.

In order to shed some light on the evaluation methods, we inspected cases in which the automatic scores and the human judgments of a realization are misaligned. Table 11 and Table 12 (concerning NLI and FR, respectively) show some samples. In Table 11, a particularly interesting example is (7): the text is extremely comprehensible for humans, however, since none of the tokens of the generated text overlaps with those of the reference, the score assigned by ROUGE-L is 0. A similar thing happens in (8) but for different reasons: SBERT is unable to get the semantic similarity between the generated text and the more natural-

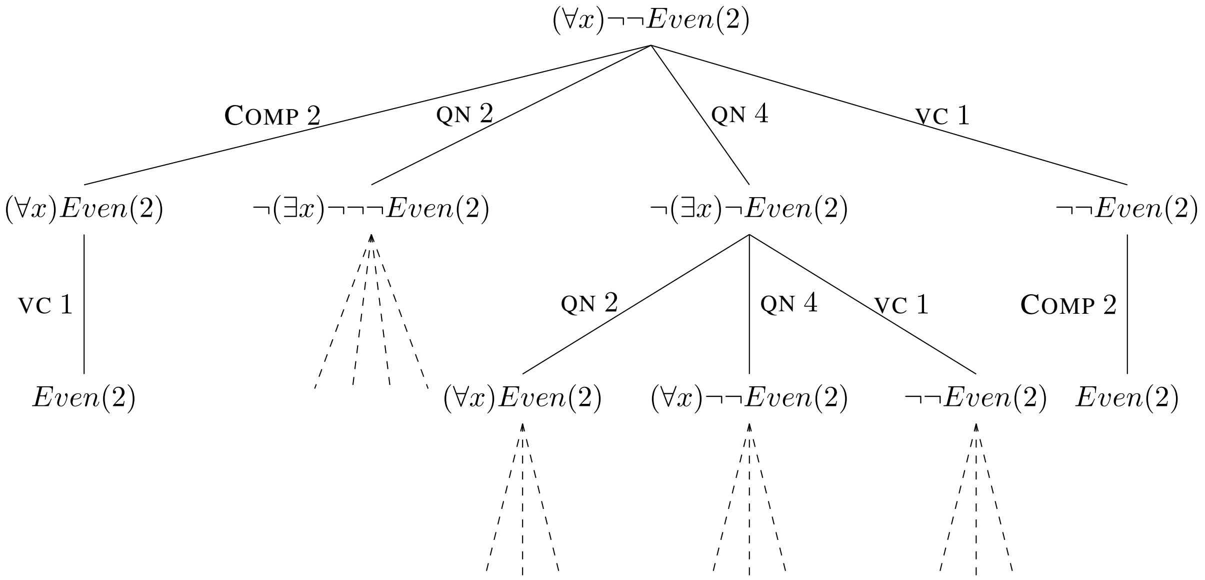


Figure 4: A part of the search tree of possible manipulations for the formula $(\forall x)\neg\neg Even(2)$, where the names of the logic laws are abbreviated (e.g., QN $1 =$ the first law of quantifier negation, see Table 8). In the system, this tree has ASTs as nodes, but for readability, the formula linearizations are displayed instead.

Linguistic Construction	Logical Construct	Example
Relative clause	Conjunction	*There exists something that is not prime.*
Adverbial clause	Conjunction	*Something is in the same row as **a**, while **a** is even.*
Adverbial clause	Implication	*Everything is small as long as it is a dodecahedron.*
		***a** is smaller than **b** if **a** is a cube.*
Correlative conjunction	Conjunction	*Both **a** is in the same row as **b** and **b** is a dodecahedron.*
Correlative conjunction	Exclusive or	*Everything is either a cube or a tetrahedron.*
Correlative conjunction	Negated disjunction	*Neither **b** nor **d** is a tetrahedron.*
Deixis	Identity	*If **b** is small, it is a tetrahedron.*
		*If **w** and **y** are tetrahedrons, they are in the same column.*
Referring expression	Identity	*If something is in front of a cube, then the cube is large.*
		*At least one of **d** and **b** is left of the other.*
Conditional mood	Implication	*If **a** had the same shape as **b**, then **c** would be in the same row as **c**.*
Modality	Implication	*If **a** is even, then there must be something that is even.*
Present participle	Implication	***a** being left of **b** implies that **b** is a dodecahedron.*
Adverbial clause	Reverse relation	***a** is to the left of **d** or the other way around.*
Modifier	Inequality	*Nothing is smaller than something else.*
Collective predicate	Distributive predication	*If **w** and **y** are tetrahedrons, they are in the same column.*

Table 9: Linguistic and stylistic constructions suggested in the post-edits of the TQA, with the logical constructs they can express or emphasize, illustrated with (slightly revised) examples.

sounding reference, thus assigning a low score to a comprehensible translation. In the opposite case (i.e., incomprehensible translations receiving high automatic scores), a noteworthy example is (9): the realization is quite convoluted (containing two negations, an implication, and the repetition of a constant), yet surprisingly, BERTScore catches the semantic similarity with the reference, even though equivalence laws are involved. Example (11) is obscure: in this case, in contrast to (8), SBERT is able to capture the semantic similarity with the reference, albeit the convoluted realization received a rather poor human judgment.

In Table 12, the realization in (12), (13), and (14) turns out to be particularly problematic. Although receiving a high score in the FR task, it is scored poorly by all the metrics, for the same reasons as above: as for ROUGE-L, the $n-$gram overlap between the realization and the reference is weak, while SBERT does not capture the semantic similarity. BERTScore is seemingly the only metric that handles semantics satisfactorily, as its score is anyhow relatively high. In the opposite case (i.e., disfluent translations receiving high automatic scores), (15) is particularly remarkable as BERTScore is capable of detecting the semantic similarity of logically equivalent constructions involving antonyms (***x*** *is smaller than* ***y*** $\equiv$ ***y*** *is larger than* ***x***). The realization in (16) is a nearly-literal translation of the original formula that is considered very disfluent by humans. Regardless, the $n-$gram overlap with the reference is prominent, so ROUGE-L gives it a relatively high score. (17) shows similar behavior to (11): SBERT surprisingly catches the semantic similarity between the realization and the reference, despite the involvement of equivalence laws.

F TRUESKILL Description

TRUESKILL was originally developed in Herbrich et al. (2006) for modeling the relative skills of players in online gaming communities, echoing Elo (1978). In higher-level terms, TRUESKILL assumes that the skill level (score) of each player (system) S_j is defined by its estimated mean performance μ_{S_j} and the uncertainty of this estimate $\sigma^2_{S_j}$. Before any match is played, μ_{S_j} is initialized to 0. These Bayesian estimates are continually updated with each match.[17] The size of the updates depends on the amount of *surprisal* and *confidence*. A player with a relatively low mean performance beating a player with a relatively high mean performance is more surprising than the opposite outcome. Thus, more surprising outcomes result in bigger updates than less surprising ones.

G TQA Questionnaire

Figure 7 presents the instruction text shown to the participants at the beginning of the survey, and Figure 8 the set of questions provided to the participants. Table 13 shows an example batch of formulae and translations used as experimental items. The filler formulae and translations present in the table were designed in such a way that the translations resembled those of Ranta (2011), they were incorrect, and their incorrectness would be easily detectable for people with a moderate amount of experience in logic.

H Human Evaluation Questionnaire

Figure 9 presents the instructions and questions shown to the participants.

[17] Note that μ_{S_j} can take negative values.

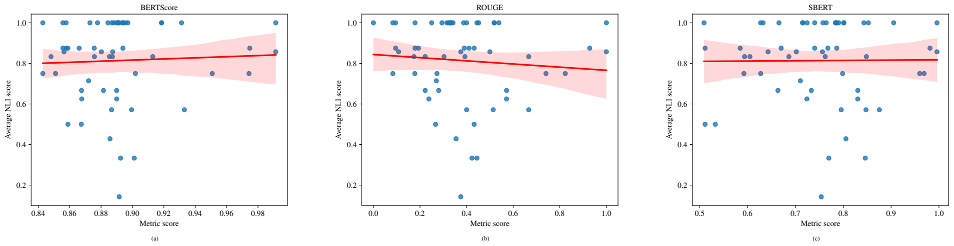


Figure 5: Scatterplots with the relationship between the average NLI score and the score assigned by the automatic metrics to the translations.

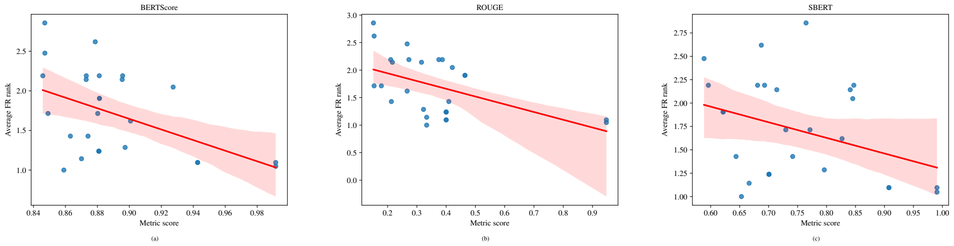


Figure 6: Scatterplots highlighting the negative correlations between the average FR rank and the score assigned by the automatic metrics to the translations.

	Formula and Reference	BASELINE	Ranta (2011)	LOLA
(1)	$\neg\neg Large(d)$ ***d*** *is large.*	*It is not the case that it is not the case that **d** is large.*	*It is not the case that **d** is not large.*	***d*** *is large.*
(2)	$\forall x \forall y((Cube(x) \wedge FrontOf(y,x)) \rightarrow Small(x))$ *If a cube has something in front of it, then it's small.*	*For all **x**, for all **y**, if **x** is a cube and **y** is in front of **x**, then **x** is small.*	*For all **x**, for all **y**, if **x** is a cube and **y** is in front of **x**, then **x** is small.*	*For all **y**, for all cubes **x**, **y** is not in front of **x** or **x** is small.*
(3)	$Smaller(f,a) \vee BackOf(f,a)$ ***f*** *is either in back of or smaller than* ***a***.	***f*** *is smaller than **a** or **f** is in back of **a**.*	***f*** *is smaller than **a** or in back of **a**.*	***f*** *is smaller than **a** or in back of **a**.*
(4)	$\exists u(Dodec(u) \wedge \neg(Large(u) \vee Small(u)))$ *Some dodecahedron is neither large nor small.*	*There is an element **u** such that **u** is a dodecahedron and it is not the case that **u** is large or **u** is small.*	*There is a dodecahedron **u** such that it is not the case that **u** is large or small.*	*It is not the case that every dodecahedron is small or large.*
(5)	$\neg\exists w(Person(w) \wedge Pet(w))$ *People are not pets.*	*It is not the case that there is an element **w** such that **w** is a person and **w** is a pet.*	*It is not the case that some person is a pet.*	*For all persons **w**, **w** is not a pet.*

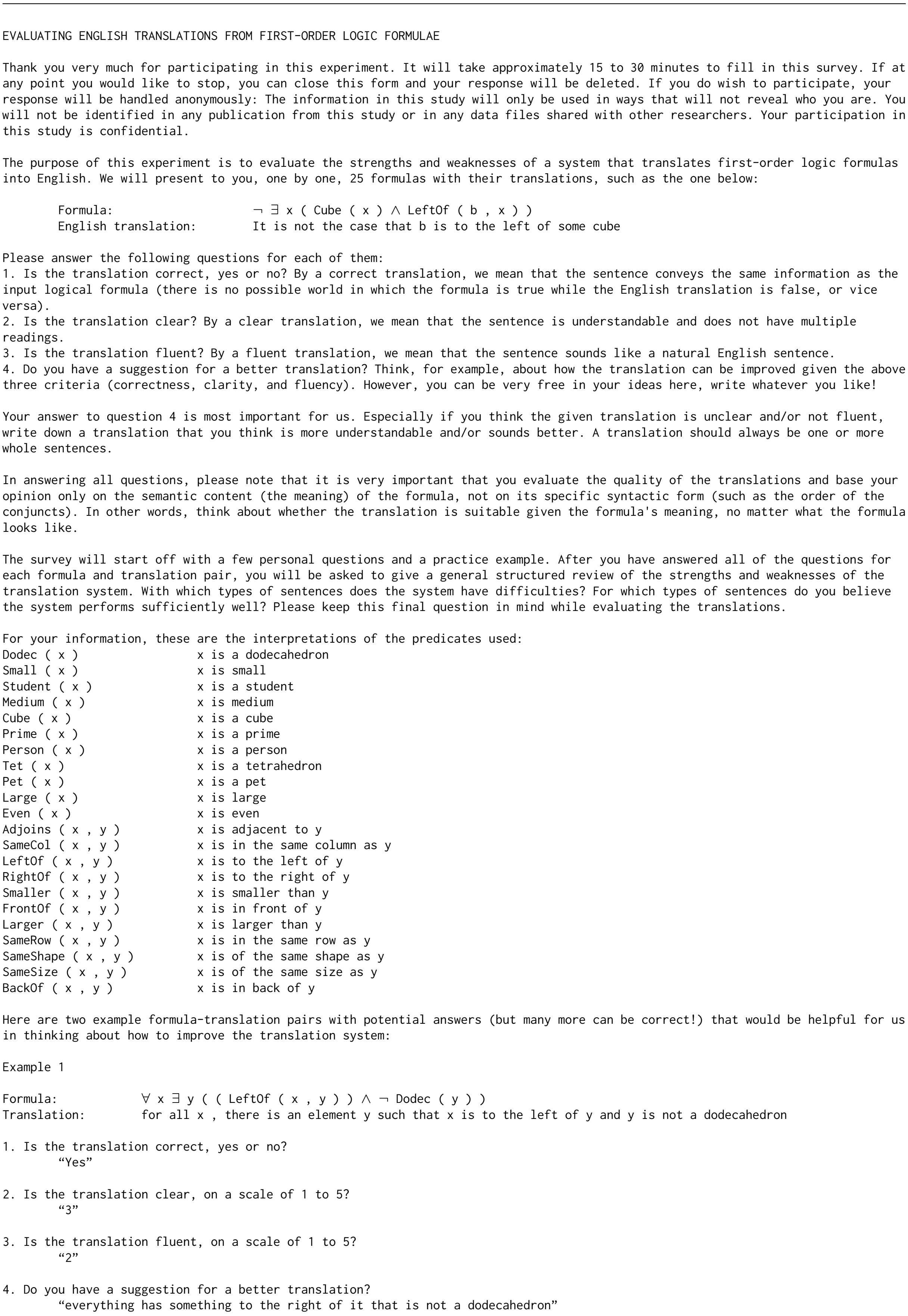
Table 10: Examples of text generated by the three systems compared in the evaluation, together with the input formula and the ground truth reference.

	Formula and Reference	Realization	System	NLI Score (↑)	Metric Score (↑)	Metric
(6)	$\forall x \neg\exists x Cube(x)$ *There are no cubes.*	*For all **x**, it is not the case that there is an element **x** such that **x** is a cube.*	Ranta (2011)	1.000	0.843	BERTScore
(7)	$\forall x \neg\exists x Cube(x)$ *There are no cubes.*	*Nothing is a cube.*	LOLA	1.000	0.000	ROUGE-L
(8)	$\exists x \exists y(Cube(x) \wedge Cube(y) \wedge Large(x) \wedge Small(y) \wedge FrontOf(x,y))$ *A large cube is in front of a small cube.*	*There is an element **x** such that there is an element **y** such that **x** is a cube and **y** is a cube and **x** is large and **y** is small and **x** is in front of **y**.*	BASELINE	1.000	0.509	SBERT
(9)	$\neg(Larger(b,a) \rightarrow \neg Larger(b,e))$ ***b*** *is larger than both **a** and **e**.*	*It is not the case that if **b** is larger than **a**, then **b** is not larger than **e**.*	Ranta (2011)	0.333	0.901	BERTScore
(10)	$\neg(Larger(b,a) \rightarrow \neg Larger(b,e))$ ***b*** *is larger than both **a** and **e**.*	*It is not the case that if **b** is larger than **a**, then **b** is not larger than **e**.*	Ranta (2011)	0.333	0.444	ROUGE-L
(11)	$\exists x(Larger(a,x) \wedge Cube(x)) \rightarrow \neg\exists y(Tet(y) \wedge \neg Smaller(a,y))$ *If **a** is larger than some cube then it is smaller than every tetrahedron.*	*If there is an element **x** such that **a** is larger than **x** and **x** is a cube, then it is not the case that there is a tetrahedron **y** such that **a** is not smaller than **y**.*	Ranta (2011)	0.333	0.846	SBERT

Table 11: Selected cases of misalignment between the normalized score retrieved from the NLI task vs. the score assigned to the realizations by some automatic metrics against the ground truth reference.

	Formula and Reference	Realization	System	FR Score (↓)	Metric Score (↑)	Metric
(12)	$\forall x((Cube(x) \wedge Small(x)) \rightarrow \exists t(Large(t) \wedge Cube(t) \wedge BackOf(x,t)))$ *Every small cube is in back of a large cube.*	*For all small cubes **x**, there is an element **t** such that **t** is large, **t** is a cube and **x** is in back of **t**.*	LOLA	1.000	0.859	BERTScore
(13)	$\forall x((Cube(x) \wedge Small(x)) \rightarrow \exists t(Large(t) \wedge Cube(t) \wedge BackOf(x,t)))$ *Every small cube is in back of a large cube.*	*For all small cubes **x**, there is an element **t** such that **t** is large, **t** is a cube and **x** is in back of **t**.*	LOLA	1.000	0.333	ROUGE-L
(14)	$\forall x((Cube(x) \wedge Small(x)) \rightarrow \exists t(Large(t) \wedge Cube(t) \wedge BackOf(x,t)))$ *Every small cube is in back of a large cube.*	*For all small cubes **x**, there is an element **t** such that **t** is large, **t** is a cube and **x** is in back of **t**.*	LOLA	1.000	0.653	SBERT
(15)	$Smaller(a,b) \wedge Smaller(e,b)$ ***b*** *is larger than both **a** and **e**.*	***a*** *is smaller than **b** and **e** is smaller than **b**.*	BASELINE	2.048	0.927	BERTScore
(16)	$\neg\exists x(LeftOf(x,a) \wedge \exists z(Smaller(x,z) \wedge LeftOf(z,b)))$ *Nothing to the left of **a** is smaller than anything to the left of **b**.*	*It is not the case that there is an element **x** such that **x** is to the left of **a** and there is an element **z** such that **x** is smaller than **z** and **z** is to the left of **b**.*	BASELINE/Ranta (2011)	1.905	0.464	ROUGE-L
(17)	$Large(a) \vee Large(c) \vee \neg(Tet(a) \wedge Tet(c))$ ***a*** *and **c** are both tetrahedra only if at least one of them is large.*	***a*** *is large or **c** is large or it is not the case that **a** is a tetrahedron and **c** is a tetrahedron.*	BASELINE	2.190	0.847	SBERT

Table 12: Selected cases of misalignment between the score (averaged) assigned to the realizations by humans in the FR task vs. the score assigned by some automatic metrics against the ground truth reference.



EVALUATING ENGLISH TRANSLATIONS FROM FIRST-ORDER LOGIC FORMULAE

Thank you very much for participating in this experiment. It will take approximately 15 to 30 minutes to fill in this survey. If at any point you would like to stop, you can close this form and your response will be deleted. If you do wish to participate, your response will be handled anonymously: The information in this study 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.

The purpose of this experiment is to evaluate the strengths and weaknesses of a system that translates first-order logic formulas into English. We will present to you, one by one, 25 formulas with their translations, such as the one below:

Formula: $\neg \exists x (Cube (x) \wedge LeftOf (b , x))$
English translation: It is not the case that b is to the left of some cube

Please answer the following questions for each of them:
1. Is the translation correct, yes or no? By a correct translation, we mean that the sentence conveys the same information as the input logical formula (there is no possible world in which the formula is true while the English translation is false, or vice versa).
2. Is the translation clear? By a clear translation, we mean that the sentence is understandable and does not have multiple readings.
3. Is the translation fluent? By a fluent translation, we mean that the sentence sounds like a natural English sentence.
4. Do you have a suggestion for a better translation? Think, for example, about how the translation can be improved given the above three criteria (correctness, clarity, and fluency). However, you can be very free in your ideas here, write whatever you like!

Your answer to question 4 is most important for us. Especially if you think the given translation is unclear and/or not fluent, write down a translation that you think is more understandable and/or sounds better. A translation should always be one or more whole sentences.

In answering all questions, please note that it is very important that you evaluate the quality of the translations and base your opinion only on the semantic content (the meaning) of the formula, not on its specific syntactic form (such as the order of the conjuncts). In other words, think about whether the translation is suitable given the formula's meaning, no matter what the formula looks like.

The survey will start off with a few personal questions and a practice example. After you have answered all of the questions for each formula and translation pair, you will be asked to give a general structured review of the strengths and weaknesses of the translation system. With which types of sentences does the system have difficulties? For which types of sentences do you believe the system performs sufficiently well? Please keep this final question in mind while evaluating the translations.

For your information, these are the interpretations of the predicates used:

Dodec (x)	x is a dodecahedron
Small (x)	x is small
Student (x)	x is a student
Medium (x)	x is medium
Cube (x)	x is a cube
Prime (x)	x is a prime
Person (x)	x is a person
Tet (x)	x is a tetrahedron
Pet (x)	x is a pet
Large (x)	x is large
Even (x)	x is even
Adjoins (x , y)	x is adjacent to y
SameCol (x , y)	x is in the same column as y
LeftOf (x , y)	x is to the left of y
RightOf (x , y)	x is to the right of y
Smaller (x , y)	x is smaller than y
FrontOf (x , y)	x is in front of y
Larger (x , y)	x is larger than y
SameRow (x , y)	x is in the same row as y
SameShape (x , y)	x is of the same shape as y
SameSize (x , y)	x is of the same size as y
BackOf (x , y)	x is in back of y

Here are two example formula-translation pairs with potential answers (but many more can be correct!) that would be helpful for us in thinking about how to improve the translation system:

Example 1

Formula: $\forall x \exists y ((LeftOf (x , y)) \wedge \neg Dodec (y))$
Translation: for all x , there is an element y such that x is to the left of y and y is not a dodecahedron

1. Is the translation correct, yes or no?
“Yes”

2. Is the translation clear, on a scale of 1 to 5?
“3”

3. Is the translation fluent, on a scale of 1 to 5?
“2”

4. Do you have a suggestion for a better translation?
“everything has something to the right of it that is not a dodecahedron”

```
Example 2

Formula:           Pet ( a ) → ∃ x Adjoins ( b , b )
Translation:       if a is a pet , then there is an element x such that x is adjacent to b

1. Is the translation correct, yes or no?
        “No”

2. Is the translation clear, on a scale of 1 to 5?
        “3”

3. Is the translation fluent, on a scale of 1 to 5?
        “1”

4. Do you have a suggestion for a better translation?
        “if a is a pet, then b is adjacent to itself”


Now it is your turn!
```

Figure 7: The instruction text shown to the participants at the beginning of the TQA.

```
0. Informed consent
    I have read the above information and understand the purpose of the research and that data will be collected from me. I also
    understand that participating in this study is completely voluntary. 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 want to withdraw from participation


1. Personal questions
    What is your gender?
            ○ Male
            ○ Female
            ○ Prefer not to say

    How old are you?

    How would you rate your knowledge of and familiarity with first-order logic? Where 1 stands for “I have been introduced to
    logic but it is long ago and I am a bit rusty”, and 5 stands for “I use logic on a daily basis”.
                    1   2   3   4   5


2. Questions for each of the formula-translation pairs in the experimental items of the batch:
    Formula:        <formula>
    Translation:    <translation>

    1. Is the translation correct? Correct means that the sentence conveys exactly the same information as the input logical
    formula.
            ○ Yes
            ○ No

    2. Is the translation clear? Clear means that the sentence is understandable and does not have multiple readings.
    (Very unclear)  1   2   3   4   5   (Very clear)

    3. Is the translation fluent? Fluent means that the sentence sounds as a natural English sentence.
    (Not fluent)    1   2   3   4   5   (Very fluent)

    4. Do you have a suggestion for a better translation? If so, then write it down here.


3. Final questions
    Give a general structured review of the strengths and weaknesses of the translation system. With which types of formulas does
    the system have difficulties? For which types of formulas do you believe the system performs sufficiently well?

    Do you have any final comments?
```

Figure 8: The set of questions provided to the participants in the TQA.

Item	Type	FOL Formula	English Translation
1	GGC	$\forall z((Cube(z) \land \exists u FrontOf(u, z)) \rightarrow Small(z))$	*for all z , if z is a cube and there is an element u such that u is in front of z , then z is small*
2	GGC	$\forall v((Dodec(v) \land \neg\exists w RightOf(w, v)) \rightarrow Small(v))$	*for all v , if v is a dodecahedron and it is not the case that there is an element w such that w is to the right of v , then v is small*
3	GGC	$\neg Cube(a) \rightarrow (Cube(c) \lor (\neg Cube(c) \rightarrow Cube(e)))$	*if a is not a cube , then at least one of these holds :* • *c is a cube* • *if c is not a cube , then e is a cube*
4	GGC	$\forall x(\forall y(Dodec(x) \land \neg RightOf(y, x)) \rightarrow Small(x))$	*for all x , if for all y , x is a dodecahedron and y is not to the right of x , then x is small*
5	GGC	$\neg\exists y(\neg Tet(y) \land \neg\exists x FrontOf(x, y))$	*it is not the case that there is an element y such that y is not a tetrahedron and it is not the case that there is an element x such that x is in front of y*
6	GGC	$\neg\exists x(\neg\exists y FrontOf(y, x) \land \neg Tet(x))$	*it is not the case that there is an element x such that it is not the case that there is an element y such that y is in front of x and x is not a tetrahedron*
7	GGC	$\forall x((Dodec(x) \land \neg\exists y RightOf(x, y)) \rightarrow \exists z LeftOf(x, z))$	*for all x , if x is a dodecahedron and it is not the case that there is an element y such that x is to the right of y , then there is an element z such that x is to the left of z*
8	GGC	$\forall y\forall x((Dodec(y) \land Tet(x)) \rightarrow FrontOf(x, y))$	*for all y , for all x , if y is a dodecahedron and x is a tetrahedron , then x is in front of y*
9	GGC	$\forall y\forall z((Cube(y) \land Dodec(z) \land BackOf(y, z)) \rightarrow Smaller(y, z))$	*for all y , for all z , if y is a cube , z is a dodecahedron and y is in back of z , then y is smaller than z*
10	GGC	$\neg(Cube(a) \land Cube(d)) \lor LeftOf(a, d) \lor LeftOf(d, a)$	*it is not the case that a is a cube and d is a cube , a is to the left of d or d is to the left of a*
11	RG	$Student(a) \lor (Medium(b) \lor \forall x SameSize(x, x))$	*a is a student , b is medium or for all x , x is of the same size as itself*
12	RG	$\forall x\neg(LeftOf(x, x) \rightarrow LeftOf(a, b))$	*for all x , it is not the case that if x is to the left of itself , then a is to the left of b*
13	RG	$Adjoins(a, b) \land ((SameRow(a, b) \land Person(b)) \rightarrow (Dodec(c) \land RightOf(c, a)))$	*all these hold :* • *a is adjacent to b* • *if a is in the same row as b and b is a person , then c is a dodecahedron and c is to the right of a*
14	RG	$\forall x RightOf(a, a)$	*for all x , a is to the right of itself*
15	RG	$\forall x\forall x\exists x SameSize(x, a)$	*for all x , for all x , there is an element x such that x is of the same size as a*
16	RG	$\exists x\forall x\neg Larger(x, x)$	*there is an element x such that for all x , x is not larger than x*
17	RG	$\neg(Adjoins(a, b) \rightarrow Adjoins(a, c)) \lor \neg(Student(c) \land Medium(a))$	*it is not the case that if a is adjacent to b , then a is adjacent to c or it is not the case that c is a student and a is medium*
18	RG	$Medium(a) \lor ((Small(b) \rightarrow Tet(b)) \rightarrow \neg Person(c))$	*at least one of these holds :* • *a is medium* • *if if b is small , then b is a tetrahedron , then c is not a person*
19	RG	$\forall x\exists x SameSize(x, a)$	*for all x , there is an element x such that x is of the same size as a*
20	RG	$(\exists x FrontOf(a, x) \rightarrow (Large(a) \land SameSize(a, b))) \lor Smaller(c, c)$	*at least one of these holds :* • *if there is an element x such that a is in front of x , then a is large and of the same size as b* • *c is smaller than itself*
21	Filler	$\neg\exists x(SameShape(a, b) \rightarrow SameRow(c, c))$	*it is not the case that there is an element x such that a is in the same shape as b and c is in the same row as itself*
22	Filler	$\forall x(Tet(x) \lor Prime(a)) \lor \exists x(Person(x) \rightarrow Student(a))$	*for all x , x is a tetrahedron or a is a prime or there is an element x such that a is a student*
23	Filler	$\exists x\forall y Larger(x, a) \land \forall y\neg Pet(b)$	*there is an element x such that for all y , x is larger than a and there is an element x such that for all y , b is not a pet*
24	Filler	$\exists x(((SameShape(x, a) \land Tet(x)) \rightarrow Adjoins(x, a))$	*for all x , if x is of the same shape as a, then x is adjacent to a*
25	Filler	$\exists x Cube(x)(Person(a) \rightarrow Adjoins(x, a))$	*there is a cube such that a is a person or x is adjacent to a*

Table 13: One example batch of formulae and translations of the experimental items used in the TQA (GGC = formulae taken from the Grade Grinder Corpus with Ranta (2011)'s translation, RG = randomly generated formulae with Ranta (2011)'s translation, Filler = randomly generated formulae with manually crafted incorrect translation).

NATURAL LANGUAGE INFERENCE & FLUENCY RANKING

Thank you very much for participating in this experiment! In this experiment, you will be performing 2 separate tasks, which will be explained to you beforehand. It will take approximately 30 minutes to complete the tasks. If at any point you would like to stop, you can close this form and your response will be deleted. If you do want to participate, your response will be handled anonymously: The information in this study 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 you wish to participate, please confirm your consent in the following question. For any questions about the survey, you can contact us.

0. Informed consent

I have read the above information and understand the purpose of the research and that data will be collected from me. I also understand that participating in this study is completely voluntary. 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 want to withdraw from participation

1. Personal questions

What is your gender?

○ Male
○ Female
○ Prefer not to say

How old are you?

How would you rate your proficiency in English?

1 2 3 4 5

How would you rate your knowledge of and familiarity with first-order logic?

1 Lower level than the ones below
2 Level of a bachelor/master student who has followed 1 or 2 classes of logic.
3 Level of a bachelor/master student who has followed more than 2 classes of logic.
4 Higher level than the ones above

From which perspective have you mainly studied logic?

○ Computational/mathematical perspective
○ Linguistic/philosophical perspective
○ Another perspective

Did you participate in our previous experiment in March 2022? This experiment was called "Evaluating English translations from First-Order Logic formulae". The participants were asked to judge the quality of English translations from First-Order Logic formulae, and provide suggestions for better translations.

○ Yes
○ No

2. Natural Language Inference

As we explained before, you will be performing 2 separate tasks in this experiment. The first one is called Natural Language Inference task, which works as follows: In each question, you are shown two sentences, which are called the premise and the hypothesis. You will be asked to think about whether the hypothesis follows from the premise or not.
For your information, the premises and hypotheses always make claims about a domain consisting of objects called A, B, C, D, E and F. There are no other objects in this domain. Some premises and hypotheses might describe weird or impossible situations. This is because the domain is part of an extraordinary world, where it can happen that something is smaller than itself, or next to itself; where something can be smaller and larger than something else at the same time; where cubes can be even and odd; where objects are not always of the same size as itself. The only thing you have to worry about, however, is whether the hypothesis is automatically true if the premise is true, no matter how odd their interpretations.

Here are two example questions to give you an idea of what the task looks like:

Example 1. Does the hypothesis automatically follow from the premise?
Premise: It is not the case that B is to the left of some cube.
Hypothesis: There is no cube.
○ Yes
○ No
○ Other (pick this option if is unclear whether the hypothesis follows from the premise, and explain why)

In this example, the correct answer is No, because the premise only states that B is not to the left of some cube, but does not state anything about the existence of cubes in general. So it does not follow from the premise that there is no cube.

Example 2. Does the hypothesis automatically follow from the premise?
Premise: For all x, x is a cube.
Hypothesis: B is a cube.
○ Yes
○ No
○ Other (pick this option if is unclear whether the hypothesis follows from the premise, and explain why)

The premise is a translation from a first-order logic formula. It quantifies over the entire domain, stating that for all objects x in the domain, x is a cube. In other words: Everything is a cube. So the correct answer is Yes, because if everything in the domain is a cube, then B, an object in the domain, is a cube.

In answering the following questions, choose the third answer option Other if it is debatable whether the hypothesis follows from the premise (e.g., if the premise is open to multiple interpretations, or if you do not understand the premise or hypothesis). Explain there shortly what is unclear. Please do not think too long about each question. If you have much trouble understanding the premise or hypothesis, choose the Other option and move on.

(Now 42 NLI questions of the following form are shown:)
Does the hypothesis automatically follow from the premise? Pick the third answer option if it is unclear whether the hypothesis follows from the premise (e.g., if the premise is open to multiple interpretations, or if you do not understand the premise or hypothesis), and explain why.
Premise: <premise>
Hypothesis: <hypothesis>
○ Yes
○ No
○ Other

3. Fluency Ranking

The purpose of this second (and final) task, which is called Fluency Ranking task, is to evaluate the fluency of English translations from first-order logic formulas. We will present to you, one by one, 20 formulas with 3 candidate translations, like in the example below:

Formula: $\neg \exists x \; (\mathrm{Cube} (x) \land \mathrm{LeftOf} (B , x))$
Translation 1: There is no element x such that x is a cube and B is to the left of x.
Translation 2: It is not the case that there is an element x such that x is a cube and B is to the left of x.
Translation 3: For all cubes x, B is not to the left of x or x is not even.

Please rank the translations by the criterion of fluency, where rank 1 stands for the most fluent, and 3 for the least fluent translation. By a fluent translation, we mean a translation that sounds as a natural English sentence. In ranking, ties are allowed. So, for example, if you think Translation 1 is best and Translation 2 and 3 are equally bad, give Translation 1 the highest rank (1), and Translation 2 and 3 the next highest rank (2), assigning nothing to the third rank.
In ranking the translations, please note that it is very important that you evaluate the fluency of the translations based only on the form of the translations (not on their adequacy given the formula). It can happen that two candidate translations are exactly the same. Please assign them the same rank always.

For your information, these are the interpretations of the predicates used in the formulas:

Dodec (x)	x is a dodecahedron
Small (x)	x is small
Student (x)	x is a student
Medium (x)	x is medium
Cube (x)	x is a cube
Prime (x)	x is a prime
Person (x)	x is a person
Tet (x)	x is a tetrahedron
Pet (x)	x is a pet
Large (x)	x is large
Even (x)	x is even
Adjoins (x , y)	x is adjacent to y
SameCol (x , y)	x is in the same column as y
LeftOf (x , y)	x is to the left of y
RightOf (x , y)	x is to the right of y

Smaller (x , y)	x is smaller than y
FrontOf (x , y)	x is in front of y
Larger (x , y)	x is larger than y
SameRow (x , y)	x is in the same row as y
SameShape (x , y)	x is of the same shape as y
SameSize (x , y)	x is of the same size as y
BackOf (x , y)	x is in back of y

(Now 20 FR questions of the following form are shown:)

Given the following formula and candidate translations, rank the translations from most fluent (1) to least fluent (3). Base your ranking only on the criterion of fluency (how natural the sentence sounds in English). Ties are allowed.

Formula:	<formula>
Translation 1:	<translation 1>
Translation 2:	<translation 2>
Translation 3:	<translation 3>

	(Most fluent) 1	2	3 (Least fluent)
Translation 1	○	○	○
Translation 2	○	○	○
Translation 3	○	○	○

4. Final question

Do you have any final comments on the survey?

Figure 9: The instructions and questions shown to the participants in the human evaluation.