Diagnosing Compositional Generalization in Transformers on ReCOGS with Compositional Graph Similarity

Bruno Leite Franco and Edson Emilio Scalabrin
Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Brazil
leite.franco@pucpr.edu.br

Abstract

This paper investigates the evaluation of compositional generalization in Transformer models on the ReCOGS benchmark. ReCOGS is a controlled synthetic benchmark designed to test whether models can systematically generalize from familiar linguistic elements to novel grammatical and semantic combinations. However, ReCOGS relies on Semantic Exact Match (SEM), a binary metric that assigns the same penalty to minor local mismatches and severe structural errors, limiting diagnostic interpretation. To address this limitation, this study introduces *Compositional Graph Similarity* (CGS), a graph-based metric that compares predicted and reference semantic structures through explicit edit operations, providing graded and interpretable structural evaluation. Controlled synthetic datasets are also used to test whether low-scoring ReCOGS categories reflect true model limitations or weaknesses in dataset coverage. Empirical results show that CGS identifies the lowest-scoring ReCOGS categories as *cp recursion* (45.0%), *obj pp to subj pp* (65.4%), and *prim to inf arg* (66.7%). Follow-up experiments show 0% SEM under depth extrapolation and constituent-role relocation, but 99.9% SEM for *prim to inf arg* in isolation. These findings support the conclusion that CGS is more informative than SEM and that Transformer limitations in ReCOGS are partly structural and partly due to dataset distribution.

1 Introduction

Compositionality is a central notion in linguistic theory and in the study of meaning representation, which states that the meaning of a complex expression arises from the meanings of its parts and from the way they are structurally combined. This property is what allows speakers to understand and produce an unbounded number of expressions, including many they have never encountered before, by systematically interpreting familiar elements in new combinations (Fodor and Pylyshyn, 1988). This principle is important for natural language processing because language understanding requires more than sensitivity to lexical co-occurrence. When evaluating Transformers and large language models, this issue has become even more relevant, as strong empirical performance often coexists with uncertainty about whether models are learning transferable compositional principles or merely exploiting statistical regularities present in training data (Lake and Baroni, 2018).

To investigate this question under controlled conditions, ReCOGS (*Revised COGS*) (Wu et al., 2023) provides a synthetic benchmark specifically designed to evaluate compositional generalization. It isolates linguistically meaningful variation in grammatical structure from unintended biases found in natural corpora. In this setting, models must generalize across systematic combinations rather than rely only on memorized surface patterns. However, the benchmark uses exact semantic correspondence between prediction and reference as its standard evaluation metric. This choice causes it to penalize qualitatively different errors equally. As a result, minor non-structural deviations and severe structural violations cannot be distinguished.

This limitation motivates the search for an evaluation metric that is better aligned with the structural nature of the task. The present work introduces Compositional Graph Similarity (CGS), a metric designed for ReCOGS that distinguishes structural from non-structural deviations and yields more interpretable diagnostic results. CGS is designed for synthetic datasets like ReCOGS, that require evaluation methods that are sensitive to logical form structure.

2 Background

2.1 ReCOGS

ReCOGS is a revised version of COGS (Kim and Linzen, 2020) designed to preserve the original

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 31–39

benchmark's semantic goals while reducing artifacts introduced by the target logical-form notation. It is constructed by applying a set of meaning-preserving changes to the original task. ReCOGS standardizes aspects of the representation by treating proper nouns more uniformly and by separating events from semantic roles in a Neo-Davidsonian style. These revisions preserve the original evaluation partition sizes while expanding the training set from 24,155 instances in COGS to 135,547 in ReCOGS (Wu et al., 2023).

The benchmark continues to target compositional generalization, but some generalization classes are especially relevant for structural analysis. In lexical-role shifts such as *Subj → Obj Proper*, *Prim → Subj Proper*, and *Prim → Obj Proper*, the model must assign a familiar item to a grammatical role not observed during training. In structural splits, the challenge is more directly compositional. In *Obj PP → Subj PP*, the model is trained on object-modifying prepositional phrases and tested on subject-modifying ones. In *CP Recursion*, test examples contain deeper clausal embedding than training examples. In *PP Recursion*, the model must interpret recursively stacked nominal modifiers of greater depth than those seen during training.

For illustration, the examples below are shown in simplified ReCOGS-style logical:

- **Obj PP → Subj PP**

  Train: Emma ate **the cake on the table**.
  `Emma(15); cake(23); table(41); eat(11) AND agent(11,15) AND theme(11,23) AND nmod.on(23,41)`

  Generalization: **The cake on the table** burned.
  `cake(23); table(41); burn(58) AND theme(58,23) AND nmod.on(23,41)`

- **CP Recursion**

  Train: Noah knew that Emma said that the cat painted.
  `Noah(26); cat(33); know(10); Emma(17) AND agent(10,26) AND ccomp(10,21) AND say(21) AND agent(21,17) AND ccomp(21,44) AND paint(44) AND agent(44,33)`

  Generalization: Noah knew that Emma said that **John saw that** the cat painted.
  `Noah(26); cat(33); know(10); Emma(17); John(28) AND agent(10,26) AND ccomp(10,21) AND say(21) AND agent(21,17) AND ccomp(21,52) AND see(52) AND agent(52,28) AND ccomp(52,44) AND paint(44) AND agent(44,33)`

- **PP Recursion**

  Train: John saw the ball in the bottle in the box.
  `ball(14); bottle(29); box(37); see(5) AND agent(5,John) AND theme(5,14) AND nmod.in(14,29) AND nmod.in(29,37)`

  Generalization: John saw the ball in the bottle in the box **on the floor**.
  `John(19); ball(14); bottle(29); box(37); floor(48); see(5) AND agent(5,19) AND theme(5,14) AND nmod.in(14,29) AND nmod.in(29,37) AND nmod.on(37,48)`

In these examples, the numbers in parentheses are variable identifiers, not numerical values or part of the lexical meaning. For instance, in `Emma(15)`, the identifier 15 denotes the entity labeled `Emma`; in `eat(11)`, the identifier 11 denotes an event labeled `eat`. Relations such as `agent(11,15)` and `theme(11,23)` then specify how these variables are connected: variable 15 is the agent of event 11, and variable 23 is its theme. The specific numbers assigned to variables are arbitrary and may be freely renamed, provided that the same identifier continues to refer to the same entity or event throughout the logical form and that distinct variables preserve the intended semantic distinctions. Thus, two logical forms that differ only by a consistent renaming of variable identifiers are semantically equivalent.

Empirically, ReCOGS remains challenging even after these revisions. In the original report, both LSTM and Transformer models still achieve near-100% performance on the IID test split, but structural generalization remains difficult. For the *Obj PP → Subj PP* split, the reported average Semantic Exact Match is 13.4% for LSTMs and 4.5% for Transformers, and performance on recursive structural splits remains in the single-digit to low-double-digit range. At the same time, lexical splits are often substantially easier, which indicates that structural recombination remains the more difficult component of the benchmark (Wu et al., 2023).

The canonical evaluation metric for ReCOGS is *Semantic Exact Match* (SEM). Under SEM, a prediction is counted as correct only if there exists a semantically consistent conversion of the variables in the predicted logical form into those of the reference. Operationally, this amounts to checking whether there is a bijection between predicted and gold variables such that corresponding variables participate in exactly the same predications; conjunctions are treated as an order-free set, and exact match is computed as set equality after variable conversion. This makes SEM more adequate than raw string matching, since it ignores irrelevant differences in conjunct order and variable names (Wu et al., 2023). However, it remains a binary metric, therefore if exact equivalence is not satisfied, all errors receive the same score. As a result, minor local deviations and severe structural failures are collapsed into the same outcome, and the metric does not transparently reveal where the semantic structure broke down. For a synthetic benchmark such as ReCOGS, whose examples may contain substantial lexical overlap while differing crucially in grammatical organization, this lack of structural granularity limits the interpretability of the evaluation.

2.2 Quality Criteria for Graph Similarity in ReCOGS

To compare graph-based semantic evaluation metrics in a principled way, this work adopts seven quality criteria for graph similarity proposed by (Opitz et al., 2020). The first is *continuous range*, meaning that the metric should map graph pairs to a continuous interval in $[0, 1]$, where 1 denotes maximal similarity and 0 minimal similarity. The second is *identity of indiscernibles*: a score of 1 should be assigned if and only if the compared graphs are identical under the intended notion of equivalence. The third is *symmetry*, so that comparing prediction to reference yields the same value as comparing reference to prediction. The fourth is *determinism*, requiring fixed inputs to always produce the same output or within a negligible error ϵ. The fifth is *absence of unjustified bias*, that is, the metric should not overvalue some graph regions or substructures in an undocumented or unwanted way.

The sixth criterion is *symbolic correspondence*, here understood as monotonicity with respect to symbolic overlap: if one graph shares more correct symbolic structure with the reference than another, its score should not be lower. The seventh is *graded correspondence*, namely the ability to assign partial credit when the prediction is close to the reference without being fully correct. In many semantic graph benchmarks, this seventh criterion is operationalized through soft lexical matching, for instance by using cosine similarity between node embeddings. However, this strategy is not appropriate for ReCOGS. The benchmark was designed with tightly controlled lexical choices and with emphasis on grammatical recombination rather than graded lexical relatedness. As a result, node labels are intended to function as discrete semantic symbols, not as points in a lexical similarity space. Introducing cosine-based partial matching would therefore inject an external notion of similarity that is not part of the benchmark's design and could blur precisely the distinction that ReCOGS is meant to test.

Under these criteria, *Semantic Exact Match* (SEM), the canonical ReCOGS metric, satisfies only part of what is desirable from a graph similarity measure. SEM only satisfies *identity*, *symmetry*, *determinism*, and *absence of unjustified bias*, because its decision rule is binary, order-independent under the allowed normalization. Nevertheless, SEM fails the *continuous range*, *symbolic correspondence* and *graded correspondence*, because it provides no partial credit and therefore cannot distinguish small local mismatches from deep structural errors.

Beyond Semantic Exact Match, other semantic graph metrics have been proposed in the literature, most notably *SMATCH* (Cai and Knight, 2013), *SemBLEU* (Song and Gildea, 2019), and *S^2MATCH* (Opitz et al., 2020). However, none of them simultaneously satisfy all seven quality criteria adopted in this work while also remaining fully appropriate for ReCOGS. According to the comparison used in this study, SMATCH satisfies continuous range, symbolic correspondence, and, up to negligible numerical variation, identity, symmetry, and determinism; it is also applicable to ReCOGS, but it does not provide graded correspondence. SemBLEU is likewise applicable to ReCOGS and has a continuous range, but it fails identity and symmetry and introduces an undesirable bias toward high-degree or central nodes, since its path-based formulation overweights errors in structurally prominent regions. S^2MATCH comes closer to a graded metric by replacing exact concept matching with soft similarity, but this graded

correspondence operates only at the lexical level. Because ReCOGS is a synthetic benchmark whose labels are intended to behave as discrete grammatical symbols rather than items in a lexical similarity space, such soft lexical matching is not appropriate in this setting. For this reason, although these metrics each capture part of what is desirable in structural evaluation, none provides the combination of formal adequacy, graded structural comparison, and ReCOGS compatibility required here.

These limitations motivate the search for semantic graph metrics that go beyond binary correctness. Existing alternatives, such as *SMATCH*, *SemBLEU*, and *S²MATCH*, partially address this need by introducing graded comparison between predicted and reference structures. However, none of them both satisfies all seven quality criteria and remains fully suitable for ReCOGS. Although they improve on SEM in certain respects, they either lack some of the formal or interpretability properties required here, or depend on lexical soft matching that is inappropriate for a synthetic benchmark whose labels are meant to operate as discrete grammatical symbols. Consequently, the need remains for a metric that supports graded structural evaluation while respecting the design principles of ReCOGS.

3 Experiment 1

This experiment evaluates whether a graph-based structural metric can provide a more informative diagnosis of compositional generalization in ReCOGS than the benchmark's canonical binary evaluation. The empirical analysis is based on predictions from the Transformer baseline reported in the original ReCOGS study. This baseline was an encoder–decoder Transformer with two layers in each component, four attention heads, learnable absolute positional embeddings, and hidden size 300 (Wu et al., 2023).

The central idea is that exact semantic match is appropriate as a strict correctness criterion, but insufficient to characterize the severity of semantic deviations. To address this limitation, the experiment introduces Compositional Graph Similarity (CGS), a continuous metric defined over semantic graphs and grounded in edit operations. The experiment asks two questions: whether CGS satisfies the quality criteria required of a graph similarity measure in the ReCOGS setting, and what additional empirical insights it reveals when applied to the predictions of the ReCOGS Transformer baseline.

3.1 Method

Let each prediction and gold reference be represented as a labeled directed multigraph

$$G = (V, E, \ell_V, \ell_E), \tag{1}$$

where V is the set of nodes, E is the multiset of directed edges, ℓ_V assigns labels to nodes, and ℓ_E assigns labels to edges. This representation is appropriate for ReCOGS because semantic correctness depends on preserving argument structure, event structure, edge direction, and role labels rather than on string identity alone. Because ReCOGS is a synthetic benchmark with controlled vocabulary and no intended graded lexical similarity, partial node matching cannot be justified through cosine similarity or embedding-space proximity. Instead, label similarity must be defined in grammatical rather than lexical terms.

Given a predicted graph G_p, a reference graph G_r, and a partial node alignment

$$\pi : V_p \to V_r \cup \{\perp\}, \tag{2}$$

an *edit listing* is defined as an ordered sequence

$$\mathcal{L}(G_p, G_r; \pi) = \langle o_1, o_2, \ldots, o_k \rangle, \tag{3}$$

where each o_i is an atomic edit operation. The operation alphabet includes node insertion, node deletion, node relabeling, edge insertion, edge deletion, edge relabeling, and edge reversal:

$$\Sigma = \left\{ \begin{array}{l} \texttt{insV}(v, \ell_V), \texttt{delV}(v), \\ \texttt{relV}(v, \ell_V \to \ell'_V), \\ \texttt{ins\_E}(u, v, \ell_E),\ \texttt{delE}(u, v, \ell_E), \\ \texttt{relE}((u, v), \ell_E \to \ell'_E), \\ \texttt{revE}(u, v, \ell_E) \end{array} \right\} \tag{4}$$

Where `insV`, `delV` and `relV` denote node insertion, deletion, and relabeling, and `insE`, `delE`, `relE` and `revE` denote edge insertion, deletion, relabeling, and reversal, respectively.

Since determining minimal graph edit distance is a NP-hard problem (Blumenthal and Gamper, 2020), the alignment π is obtained heuristically in two stages by applying the Hungarian algorithm (Kuhn, 1955).

To define graded penalties for relabeling, let $\mathcal{T} = (\mathcal{C}, \prec)$ be a rooted grammatical taxonomy (Figure 1 was the taxonomy used for this study).

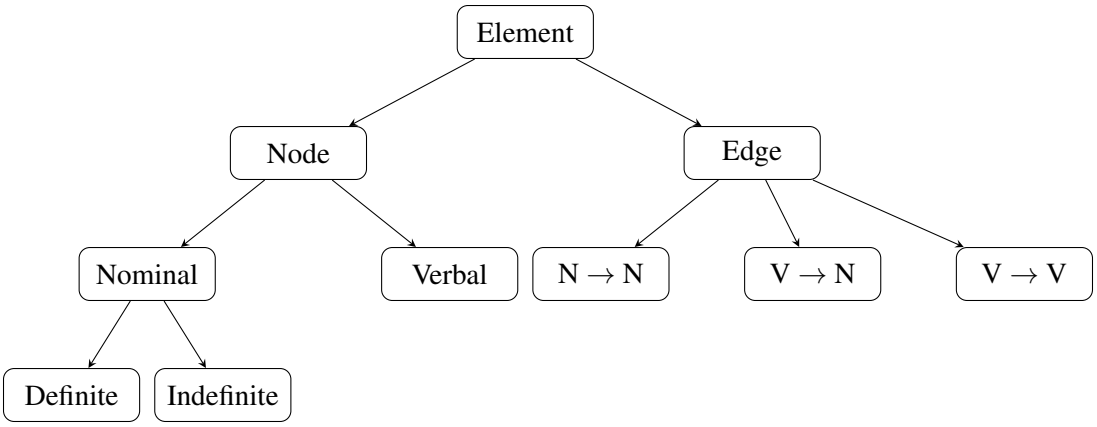

Figure 1: Taxonomy of grammatical classes for edit cost.

Node labels and edge labels are mapped to grammatical classes through a function:

$$\mathrm{cls}(\ell) \in \mathcal{C}. \tag{5}$$

Let $\mathrm{d}(c)$ be the depth of class c in the taxonomy, and let $\mathrm{a}(c_1, c_2)$ denote the lowest common ancestor of c_1 and c_2 and x and y denote class of ℓ and ℓ' respectively. The grammatical distance between labels is then defined as:

$$d_{\mathcal{T}}(\ell, \ell') = \frac{d(x) + d(y) - 2\,d(a(x,y)) + 2}{d(x) + d(y) + 2} \tag{6}$$

This definition ensures that relabelings within the same grammatical branch are penalized less than relabelings across branches, which is crucial in ReCOGS, where grammatical function matters but lexical similarity is not part of the task design.

The total grammatical edit cost is defined as

$$\mathrm{Cost}_{\mathrm{gram}}(\mathcal{L}) = \sum_{o \in \mathcal{L}} \mathrm{cost}_{\mathrm{gram}}(o), \tag{7}$$

with operation costs

$$\begin{aligned}
\mathrm{cost}_{\mathrm{gram}}(\texttt{relV}(v, \ell_V \to \ell'_V)) &= d_{\mathcal{T}}(\ell_V, \ell'_V),\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{relE}((u,v), \ell_E \to \ell'_E)) &= d_{\mathcal{T}}(\ell_E, \ell'_E),\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{revE}(u, v, \ell_E)) &= \rho_{\mathrm{rev}},\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{insE}(u, v, \ell_E)) &= 1,\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{delE}(u, v, \ell_E)) &= 1,\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{insV}(v, \ell_V)) &= 1,\\
\mathrm{cost}_{\mathrm{gram}}(\texttt{delV}(v)) &= 1.
\end{aligned} \tag{8}$$

In the present formulation, edge reversal is assigned a fixed intermediate penalty ρ_{rev}, reflecting the fact that direction errors are more severe than a small within-branch relabeling but less severe than adding or deleting structure altogether.

To make scores comparable across instances of different sizes, the cost is normalized by a symmetric size term:

$$L_{\mathrm{sym}} = \min\big(|V_r| + |E_r|,\ |V_p| + |E_p|\big), \tag{9}$$

$$\mathrm{Cost}^{\mathrm{norm}}_{\mathrm{gram}}(G_p, G_r; \pi) = \frac{\mathrm{Cost}_{\mathrm{gram}}(\mathcal{L}(G_p, G_r; \pi))}{\max(L_{\mathrm{sym}}, 1)}. \tag{10}$$

Finally, *Compositional Graph Similarity* is defined as

$$\mathrm{CGS}(G_p, G_r; \pi) = \exp\big(-\mathrm{Cost}^{\mathrm{norm}}_{\mathrm{gram}}(G_p, G_r; \pi)\big). \tag{11}$$

By construction,

$$\mathrm{CGS}(G_p, G_r; \pi) \in (0, 1], \tag{12}$$

with value 1 if and only if the normalized edit cost is 0, and values decreasing continuously as structural divergence increases.

Under this definition, CGS satisfies the seven quality criteria required for graph similarity in the ReCOGS setting. First, it has a continuous range in $(0, 1]$. Second, it satisfies identity of indiscernibles, since $\mathrm{CGS} = 1$ holds only when no edit is required. Third, it is symmetric, given symmetric operation costs. Fourth, it is deterministic once alignment and tie resolution are fixed. Fifth, it avoids unjustified bias because penalties are explicitly documented and shared across graph regions rather than emerging from hidden path-counting effects. Sixth, it satisfies symbolic correspondence: if a prediction shares more correct graph structure with the reference, the number and severity of edits necessarily decrease, which increases CGS. Seventh, it

Criterion	Exact Match	SMATCH	SemBLEU	S^2MATCH	CGS (Ours)
Continuous Range	✗	✓	✓	✓	✓
Identity	✓	$\checkmark_\epsilon$	✗	$\checkmark_\epsilon$	$\checkmark_\epsilon$
Symmetry	✓	$\checkmark_\epsilon$	✗	$\checkmark_\epsilon$	$\checkmark_\epsilon$
Determinacy	✓	$\checkmark_\epsilon$	✓	$\checkmark_\epsilon$	$\checkmark_\epsilon$
Low bias	✓	✓	✗	✓	✓
Symbolic Matching	✗	✓	✗	✓	✓
Graded Matching	✗	✗	✗	$\checkmark_{\text{LEX}}$	$\checkmark_{\text{GRA}}$
Applicable to ReCOGS	✓	✓	✓	✗	✓

$\checkmark_\epsilon$: Satisfies up to statistically negligible numerical error;
$\checkmark_{\text{LEX}}$: Graded matching only at the lexical level.
$\checkmark_{\text{GRA}}$: Graded matching only at the gramatical level.

Table 1: Comparison by criteria (rows) and metrics (columns) for use in ReCOGS.

supports graded correspondence through taxonomically grounded relabel penalties and structural edit costs.

3.2 Result

As presented in Table 1, CGS is an adequate semantic evaluation metric for ReCOGS. Unlike Semantic Exact Match, CGS satisfies all seven quality criteria adopted in this work. At the same time, it remains applicable in a benchmark such as ReCOGS, where node labels are not partially matched through lexical similarity.

When applied to ReCOGS predictions, CGS highlighted three test categories with particularly low structural similarity: *cp recursion* (CGS $=$ 45.0%), *obj pp to subj pp* (CGS $=$ 65.4%), and *prim to inf arg* (CGS $=$ 66.7%). This makes these categories especially informative for further investigation.

Rather than treating these low scores only as endpoint evaluation results, they were used as signals for the next stage of experimentation. The purpose of these follow-up datasets was to determine whether the observed failures were caused by the intrinsic difficulty of the target generalization itself, or instead by insufficient exposure to the relevant structural pattern in the original ReCOGS training set.

4 Experiment 2

Experiment 1 showed that the three lowest-scoring ReCOGS categories under CGS were *cp recursion*, *prim to inf arg*, and *obj pp to subj pp*. Because these categories were the ones in which structural similarity also degraded most strongly, they were selected for a second experiment based on synthetic controlled datasets. The purpose of Experiment 2 was not merely to test if these low scores are due to a model weakness or insufficient structural representation of the relevant pattern in the original training distribution. To make this distinction explicit, each of the three phenomena was tested in isolation, with datasets designed so that the target generalization was the only source of out-of-distribution pressure.

4.1 Method

Three synthetic diagnostic datasets were constructed, each corresponding to one of the categories with the lowest CGS in Experiment 1. The first targeted *cp recursion* involved training 5 models by varying the maximum supervised depth of clausal embedding (from 1 to 5) and then evaluating the model on test sets with depths ranging from 0 to 5 (even if the model was not exposed to that depth during training). The second targeted *prim to inf arg*, in which the model was trained on isolated lexical items and on sentences following the pattern `A V_1 to V_2`, was subsequently tested on examples in which those lexical items were integrated into full sentential structures.The third targeted *obj pp to subj pp* and involved the creation of seven models, each trained on a different non-empty combination of the three argument positions that could host a structurally complex nominal constituent, so that the effect of relocating such complexity across positions could be evaluated in isolation.

The synthetic datasets were generated with a template-based sentence generator designed to produce paired surface sentences and ReCOGS-style logical forms. The generator samples lexical items from WordNet (Miller, 1992). Verbs are sampled in base form and inflected to simple past for the

surface sentence, while the corresponding logical form uses the base verb. Nouns are sampled as lowercase single-token lemmas, and prepositional modifiers are generated using the prepositions *in*, *on*, and *beside*.

The generator uses controlled templates so that the source of generalization pressure can be isolated. For the primitive-to-infinitival-argument condition, the basic template follows the form $A\ V_1\ T$ *to* R, with logical forms encoding the corresponding agent, theme, and recipient relations. For the constituent-relocation condition, one nominal argument is expanded into a complex noun phrase of the form $N\ P\ C$, where P is a preposition and C is a complement noun. The complex constituent can appear in the agent, theme, or recipient position, yielding separate datasets for each possible host role. Training and test conditions are then defined by controlling which argument positions contain complex constituents during training and which positions are held out for generalization.

The construction procedure also enforces distributional controls intended to prevent trivial lexical explanations for the results. Within each generated sentence, role fillers are kept distinct, lexical items are balanced across grammatical roles as far as possible, and verb, noun, and preposition usage is made approximately uniform. The generator further avoids repeated local pairings when possible, using randomized construction with limited backtracking and only minimal relaxation of pair-uniqueness constraints when necessary. As a result, the diagnostic datasets differ primarily in the structural configuration being tested rather than in uncontrolled lexical frequency effects.

4.2 Result

As summarized in Table 2, the comparison between ReCOGS and the diagnostic datasets reveals that the three low-scoring categories selected from Experiment 1 do not reflect a single underlying source of failure. Whereas *cp recursion* and *obj pp to subj pp* remain difficult under controlled isolation, *prim to inf arg* shifts from failure in ReCOGS to near-ceiling performance in the diagnostic setting.

In the *cp recursion* dataset, models remained highly accurate within the depth regime represented in training, but failed categorically under depth extrapolation. More specifically, inlier performance was above 99% for all in-support conditions from P2 onward, whereas the shallowest inlier conditions were noticeably lower, with 78.1% Exact Match for the P0 model on test P0 and 75.4% for the P1 model on test P1. By contrast, whenever test depth exceeded the maximum depth observed during training, Semantic Exact Match dropped to 0%, independently of the model's training regime. This pattern indicates that the models learned depth-specific configurations rather than an abstract recursive rule that could generalize beyond the structural support available in training.

Category	**ReCOGS**	**DDG**
cp recursion	12.7%	0.0%
prim to inf arg	0.0%	99.9%
obj pp to subj pp	11.9%	0.0%

Table 2: Exact Match comparison between ReCOGS and our Diagnostic Datasets Generalization (DDG).

The sub-ceiling results in the shallow inlier conditions are best interpreted as a side effect of the random node-indexing scheme introduced to enable extrapolation to larger logical forms. In low-depth examples, where the number of nodes is small, the model appears to face greater difficulty abstracting away from index variability and isolating the underlying compositional structure. As a result, the approximately 0.75 inlier accuracies observed for P0 and P1 likely reflect noise induced by random indices in structurally small graphs, rather than failure on the recursive phenomenon itself.

The result for *prim to inf arg* was markedly different. In the isolated synthetic test, the model was highly successful (99.9% Semantic Exact Match), which shows that this generalization can, in fact, be learned reliably when the relevant structure is sufficiently represented during training. The discrepency of this high values with those in the ReCOGS dataset inplies a dataset weakness in the original dataset since only 6.2% of the origial dataset has the *xcomp* relationship or weakly cyclical logical forms (the rest of the dataset has acyclical logical forms).

The *obj pp to subj pp* experiment yielded a pattern closely parallel to that of *cp recursion*. In all inlier conditions, Semantic Exact Match remained above 99.8%, indicating that the models could reliably handle structurally complex noun phrases as long as those phrases appeared in the thematic positions supported during training. In contrast, performance dropped to 0% in every outlier condition, that is, whenever a structurally complex nominal constituent had to be reanchored in a thematic

position not represented as complex in the corresponding training distribution. This sharp contrast shows that the difficulty is not due merely to lexical novelty or to the presence of internal nominal modification. Rather, it arises from the relocation itself: the model fails when a constituent enriched by internal structure must be reassigned to a new grammatical role while preserving the correct verb–argument dependencies.

Taken together, these results show that low CGS can arise from different underlying causes. In *cp recursion* and *obj pp to subj pp*, the isolated experiments preserved the original pattern of difficulty, indicating genuine structural limitations in hierarchical embedding and constituent reanchoring. In *prim to inf arg*, however, the isolated experiment produced high performance, suggesting that the original failure in ReCOGS is largely due to insufficient structural representation of the relevant configuration in the training data.

5 Conclusion

This work argued that *Compositional Graph Similarity* (CGS) is a more adequate metric than *Semantic Exact Match* for evaluating compositional generalization in ReCOGS. Whereas Semantic Exact Match is restricted to a binary notion of correctness, CGS provides a graded structural comparison, satisfies the seven quality criteria adopted in this work, and remains applicable in synthetic semantic datasets where lexical soft matching is not appropriate. Because CGS is derived from explicit edit operations, it also offers a more interpretable account of model behavior, making it possible to distinguish minor local deviations from genuine structural failures.

The experiments further showed that low performance should not always be interpreted in the same way. In the *prim to inf arg* case, the isolated synthetic test yielded high accuracy, suggesting that the poor result in ReCOGS is not evidence of an inability to perform the target generalization itself, but rather of a limitation in the dataset distribution. More specifically, the relevant structural configuration appears to be underrepresented in the original benchmark. This result indicates that Transformer models are in fact capable of abstracting primitive lexical items and preserving global semantic structure when the required pattern is adequately represented during training.

By contrast, the results for *obj pp to subj pp* and *cp recursion* point to a different conclusion. In both cases, even controlled synthetic testing preserved the pattern of failure, indicating that the problem is not merely a property of the original dataset distribution. Instead, these results suggest that the models fail to abstract reusable substructures in a sufficiently systematic way. As a consequence, they struggle when required to interact with a novel structure, even when that structure is composed only of substructures that were individually familiar during training.

More broadly, this work suggests that compositional generalization should be evaluated not only by whether a final answer is correct, but also by how the predicted semantic or inferential structure differs from the target structure. In natural language inference, prior work has shown that high aggregate accuracy can mask reliance on shallow syntactic heuristics, such as lexical overlap or subsequence matching (McCoy et al., 2019). Similarly, recent reasoning benchmarks increasingly emphasize structured logical or proof-based representations, as in ProofWriter (Tafjord et al., 2021) and FOLIO (Han et al., 2024). For such tasks, a graph-based evaluation framework could complement exact-match or label-based accuracy by identifying whether an error reflects a minor local mismatch, an incorrect role assignment, a broken inference step, or a deeper failure to preserve compositional structure. Thus, the contribution of CGS is not only a metric for ReCOGS, but also an example of how future evaluations of semantic parsing, NLI, and reasoning systems can move toward more diagnostic, structure-sensitive assessment.

6 Limitations

CGS depends on design choices in graph construction, alignment, and edit costs. Although these choices were made to reflect grammatical structure and to satisfy the quality criteria adopted in this work, different taxonomies or cost schemes could lead to somewhat different similarity profiles.

To assess if the observed failures to abstract reusable substructures from transformer models are not limited to the ReCOGS dataset. Future work should apply the same combination of structural evaluation and controlled synthetic isolation to other benchmarks.

References

David B. Blumenthal and Johann Gamper. 2020. On the exact computation of the graph edit distance. *Pattern Recognition Letters*, 134:46–57. Applications of Graph-based Techniques to Pattern Recognition.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

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

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, Florence, Italy. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation. *Transactions of the Association for Computational Linguistics*, 11:1719–1733.