Evaluating Structural and Linguistic Quality in Urdu DRS Parsing and Generation through Bidirectional Evaluation

Muhammad Saad Amin, Luca Anselma and Alessandro Mazzei
Department of Computer Science, University of Turin, Italy
{muhammadsaad.amin, luca.anselma, alessandro.mazzei}@unito.it

Abstract

Evaluating Discourse Representation Structure (DRS)-based systems for semantic parsing (Text-to-DRS) and generation (DRS-to-Text) poses unique challenges, particularly in low-resource languages like Urdu. Traditional metrics often fall short, focusing either on structural accuracy or linguistic quality, but rarely capturing both. To address this limitation, we introduce two complementary evaluation methodologies—Parse-Generate (PARS-GEN) and Generate-Parse (GEN-PARS)—designed for a more comprehensive assessment of DRS-based systems. PARS-GEN evaluates the parsing process by converting DRS outputs back to the text, revealing linguistic nuances often missed by structure-focused metrics like SMATCH. In contrast, GEN-PARS assesses text generation by converting generated text into DRS, providing a semantic perspective that complements surface-level metrics such as BLEU, METEOR, and BERTScore. Using the Parallel Meaning Bank (PMB) dataset, we demonstrate our methodology in Urdu, uncovering unique insights into the structural and linguistic interplay of Urdu. The findings show that traditional metrics frequently overlook the complexity of linguistic and semantic fidelity, especially in low-resource languages. Our dual approach offers a robust framework for evaluating DRS-based systems, improving semantic parsing and text generation quality[1].

1 Introduction

DRS is central to advanced semantic processing, providing a flexible and language-neutral framework for capturing complex semantic nuances beyond basic text interpretation (Kamp and Reyle, 1993), including phenomena such as negation and quantification (Kamp and Reyle, 2013; Jaszczolt and Jaszczolt, 2023). Its adaptability makes DRS ideal for multilingual natural language processing (NLP) systems, offering a unified way of representing meaning across languages with diverse structural and syntactic properties (Bos, 2023).

DRS parsing (van Noord et al., 2018; Noord, 2019; van Noord et al., 2019) and generation (Wang et al., 2021; Amin et al., 2022; Liu et al., 2021; Amin et al., 2024) are reversible processes which pose unique challenges, especially when working with Urdu—a morphologically rich language. Urdu exhibits different syntactic structures and semantic expressions, making accurate evaluation difficult due to the limitations of traditional structural and surface-level metrics (Butt and King, 2002; Bögel et al., 2009). Existing evaluations often fail to fully account for linguistic and structural accuracy across languages, which is essential for ensuring meaningful cross-linguistic semantic representation. This gap has motivated our development of innovative evaluation methods to bridge structural precision with linguistic adequacy in DRS-based systems.

Our research primarily aims to create evaluation frameworks that integrate both *structural* and *linguistic* (in the sense of *surface-level*) assessments. To accomplish this, we introduce two bidirectional evaluation paradigms—PARS/PARS-GEN and GEN/GEN-PARS. The former assesses parsing quality by examining the linguistic coherence of the text generated from DRS structures, moving beyond traditional metrics to provide insights into how well structural accuracy supports meaningful language representation. Conversely, the latter evaluates generation quality by analyzing the semantic consistency of parsed structures derived from generated text, offering a deeper perspective than surface-level comparisons alone.

Semantic parsing evaluation typically relies on structural metrics like SMATCH (Cai and Knight, 2013), which assesses roles or concepts-based overlaps between predicted and reference DRS

[1] https://github.com/saadamin2k13/counter-evaluations-for-urdu.

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 33–43
January 20, 2025.

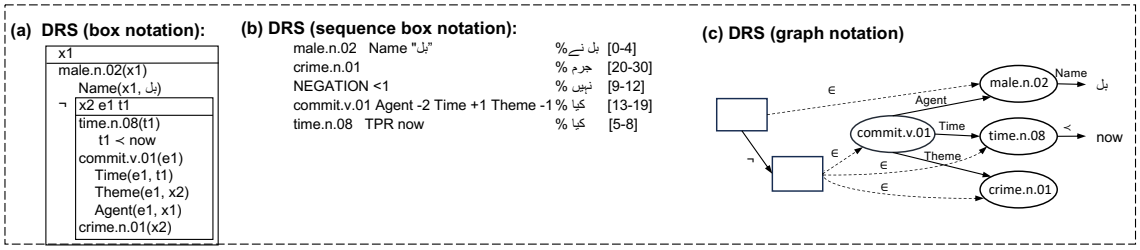

Figure 1: Different graphical representations of DRS for the text "Bill didn't commit the crime."

graphs (Kamp et al., 2010). While valuable for evaluating structural accuracy, this metric often misses essential linguistic subtleties and penalizes the overall evaluation. For instance, two DRS representations with minor structural divergences, such as `Quantity` and `Index`, obtained a significantly low SMATCH score despite near-identical semantics (Ex. 4, Table 1). Such distinctions illustrate how structural metrics alone may fall short in capturing the semantic nuances, coherence, and pragmatic meaning crucial to linguistic representation. This limitation inspired the development of the PARS/PARS-GEN approach, which leverages text generation to assess parsing quality, highlighting linguistic phenomena that structural metrics might otherwise overlook.

Text generation from DRS also poses a unique evaluation challenge. Traditional metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and even recent metrics like BERTScore (Hanna and Bojar, 2021) prioritize surface-level similarities between generated and reference texts. However, given the diversity of natural language, there can be multiple valid expressions for the same meaning. For example, the sentences ("*John gave Mary some money*") and ("*John gave the money to Mary*") convey similar meanings, but their syntactic variations lead to low scores under traditional evaluations, despite perfect semantic equivalence in DRS representation (Ex. 4, Table 2). To address this, we propose the GEN/GEN-PARS paradigm, which evaluates generated text by parsing it back into DRS, offering a structural evaluation perspective that complements surface-level metrics.

In this context, our research investigates several critical questions: (i) How can the evaluation of semantic parsing and text generation be improved beyond existing structural and surface-level metrics? (ii) How does structural accuracy in semantic parsing influence linguistic quality in text generation? (iii) How can surface-level evaluations be enhanced by assessing individual lexical entities? (iv) Can the reversible nature of semantic parsing and text generation be exploited for improved evaluations? and (v) Do these alternate evaluations correlate with each other and are they statistically significant?

To address these questions, this paper makes the following key contributions: (i) it introduces novel evaluation paradigms, PARS/PARS-GEN and GEN/GEN-PARS, which reveal unique insights into the language's syntactic variability and complex semantic structures that traditional metrics often overlook; (ii) the PARS/PARS-GEN paradigm uses linearized text to mitigate non-optimal outcomes in SMATCH's greedy search algorithm, enabling a more intuitive and human-centered approach to parsing evaluation; (iii) through the GEN/GEN-PARS evaluation, it identifies semantic and syntactic issues at a node level, examining lexical DRS concepts like nouns, verbs, adjectives, and adverbs within the generated DRS to provide a granular view of the generation quality, ultimately facilitating a balanced metric that captures both structural and linguistic fidelity; and (iv) it proposes a detailed Pearson correlation analysis between PARS/PARS-GEN, GEN/GEN-PARS. The observed statistically significant correlations underscore the robustness of our approach and demonstrate the effectiveness of combining structural and linguistic assessments in DRS-based semantic processing. Figure 1 contains different graphical representations of the DRS containing: (a) box format; (b) variable-free format; and (c) graph notation of the DRS. For our experimentation, we used the variable-free representation of the DRS (Figure. 1(b)) in its linearized format, as it is compatible with the sequence-to-sequence models. Additionally, we utilized its graph notation (Figure. 1(c)) to evaluate semantic parsing using SMATCH.

Ex. No	Gold Text	Gold (DRS)	PARS (DRS)	PARS (SMATCH)
1	ٹام نے ایک نیا پک اپ خریدا۔ (“Tom bought a new pickup.”)	male.n.02 Name “ٹام” new.a.05 AttributeOf +1 pickup.n.01 buy.v.01 Agent -3 Theme -1 Time +1 time.n.08 TPR now	male.n.02 Name “ٹام” new.a.01 AttributeOf +1 pick_up.n.01 buy.v.01 Agent -4 Beneficiary -3 Theme -1 Time +1 time.n.08 TPR now	00.00
2	ٹام مریم کو اپنے کتے کی تصویر دکھاتا ہے۔ (“Tom shows Mary a picture of his dog.”)	male.n.02 Name “ٹام” female.n.02 Name “مریم” male.n.02 ANA -2 dog.n.01 Owner -1 picture.n.01 Topic -1 show.v.04 Agent -5 Recipient -4 Theme -1 Time +1 time.n.08 TPR now	male.n.02 Name “ٹام” female.n.02 Name “مریم” female.n.02 ANA -1 dog.n.01 Owner -1 photo.n.01 Creator -1 show.v.01 Agent -4 Recipient -3 Recipient -1 Time +1 time.n.08 TPR now	69.23
3	آج میری گردن میں درد ہے۔ (“Today I have a pain in my neck.”)	day.n.03 TCT now time.n.08 TIN -1 person.n.01 EQU speaker neck.n.01 pain.n.01 Location -1 have.v.16 Time -4 Experiencer -3 Stimulus -1 Time +1 time.n.08 EQU now	person.n.01 EQU speaker neck.n.01 hurt.v.01 Patient -2 Patient -1 Time +1 time.n.08 EQU now	60.00
4	تیرہ افراد کو گرفتار کر لیا گیا۔ (“Thirteen people were arrested.”)	quantity.n.01 EQU 13 person.n.01 Quantity -1 arrest.v.01 Patient -1 Time +1 time.n.08 TPR now	quantity.n.01 EQU 30 person.n.01 Quantity -1 arrest.v.01 Patient -3 Time +1 time.n.08 TPR now	00.00
5	میرے پاس بہت پیسہ ہے۔ (“I have a lot of money.”)	person.n.01 EQU speaker money.n.01 Quantity + get.v.01 Pivot -2 Theme -1 Time +1 time.n.08 TPR now	person.n.01 EQU speaker have.v.01 Pivot -1 Theme +2 Time +3 quantity.n.01 EQU +1 quantity.n.01 EQU +1 money.n.01 Quantity + time.n.08 EQU now	57.89

Table 1: Structural overlap-based evaluation measures: highlighting limitations of SMATCH. English translations are mentioned in brackets. PARS scores are in %.

The remaining sections are organized as follows: Section 2 discusses the limitations of current evaluation approaches in detail. Section 3 presents our novel evaluation methodologies and describes the experimental setup and implementation details. Section 4 presents reversible evaluation measures and correlation analysis. Finally, Section 5 concludes with limitations.

2 Limitations in Current Evaluations

The evaluation of semantic parsing and text generation system presents unique challenges that conventional metrics often struggle to address comprehensively. This section examines these limitations in detail and establishes the motivation for our proposed evaluation approaches.

Parsing Limitations: Traditional evaluation metrics, such as SMATCH (Cai and Knight, 2013), SMATCH++ (Opitz, 2023), and SemBLEU (Song and Gildea, 2019), focus on assessing structural similarities between predicted and reference DRS. SMATCH, for instance, employs a greedy hill-climbing algorithm that matches nodes across logical structures. This approach, however, often results in suboptimal evaluations, especially in cases where structural differences do not reflect actual semantic deviations. For example, SMATCH assigns a zero score to the DRS representation for (“*Tom bought a new pickup*”), despite the semantic content being essentially equivalent in both gold and predicted DRS. The low-score is due to minor structural differences, underscoring a limitation of SMATCH’s focus on structural alignment rather than semantic equivalence (Ex. 1, Table 1).

Additionally, SMATCH’s handling of semantic relationships is limited, as it treats DRS nodes as isolated entities. This limitation is evident in Ex. 2 (“*Tom shows Mary a picture of his dog*”), where differences in role modifiers like (“`Topic`” and “`Creator`”) for “picture” results in a SMATCH score of 69.23. The metric’s penalty for these isolated structural variations, without accounting for the underlying semantic alignment, highlights its tendency to overlook contextually equivalent expressions when modifiers are altered or substituted. This penalization is further illustrated in Ex. 3 (“*Today I have a pain in my neck*”), where SMATCH deducts points based on minor discrepancies in the verb sense, yielding a score of 60.00 despite the overall message being well-preserved across both DRS.

In Ex. 4 (“*Thirteen people were arrested*”), SMATCH once again assigns a score of zero, this time due to an inconsistency in the numerical value between gold (13) and predicted (30) DRS. This significant deduction overlooks that the core event—people being arrested—is accurately conveyed. Ex. 5 (“*I have a lot of money*”) further emphasizes SMATCH’s limitations, where minor numerical and role-discrepancies lead to a score of 57.89, despite the intended meaning being largely retained. These examples collectively underscore that SMATCH’s sensitivity to structural changes can cause unfairly low scores even when semantic content is mostly preserved.

Ex. No.	Gold DRS	Gold Text	GEN Text	GEN Scores				
				BLEU	METEOR	ROUGE	chrF	B_Scr.
1	person.n.01 EQU speaker ashamed.a.01 Experiencer -1 Time +1 NEGATION <1 time.n.08 EQU now	میں شرمندہ نہیں ہوں۔ ("I am not ashamed.")	میں ابھی تک شرمندہ نہیں ہوں۔ ("I'm not shy yet.")	16.67	11.90	19.99	21.95	78.96
2	person.n.01 EQU speaker quantity.n.01 EQU 24 pencil.n.01 Quantity -1 buy.v.01 Agent -3 Theme -1 Time +1 time.n.08 TPR now	میں نے دو درجن پنسلیں خریدیں۔ ("I bought two dozen pencils.")	میں نے 24 پنسلیں خریدیں۔ ("I bought 24 pencils.")	49.12	43.31	54.54	39.79	92.09
3	NEGATION <1 person.n.01 NEGATION <1 leave.v.01 Theme -1 Time +1 time.n.08 TPR now	سب چلے گئے۔ ("Everyone left.")	اب سب چلے گئے ہیں۔ ("All have now left.")	50.00	32.25	57.14	29.38	88.74
4	male.n.02 Name "جان" female.n.02 Name "مریم" money.n.01 give.v.03 Agent -3 Recipient -2 Theme -1 Time +1 time.n.08 TPR now	جان نے مریم کو کچھ پیسے دیے۔ ("John gave Mary some money.")	جان نے رقم مریم کو دی۔ ("John gave the money to Mary.")	56.43	57.52	61.54	48.84	89.99
5	male.n.02 time.n.08 YearOfCentury '198X' bear.v.02 Patient -2 Time -1 Time +1 time.n.08 TPR now	وہ اسی کی دہائی میں پیدا ہوئے۔ ("He was born in the eighties.")	وہ اسّی میں پیدا ہوئے۔ ("He was born in 198X.")	42.32	37.04	44.15	34.12	79.26

Table 2: Semantic overlap-based evaluation measures: highlighting limitations of automatic evaluation metrics for text generation. Note: B_Scr. = BERTScore.

To address these limitations, our PARS-GEN approach rephrases DRS outputs as natural language text, enabling the use of complementary evaluation metrics like chrF, METEOR, and BERTScore, which emphasizes semantic accuracy. By generating interpretable text from DRS, PARS-GEN provides a holistic evaluation of parsing quality and captures linguistic nuances that structural metrics like SMATCH often miss. Through this approach, we enhance the accessibility and interpretability of semantic fidelity assessment, ensuring a more accurate and inclusive evaluation across diverse language structures and semantics.

Generation Limitations: Traditional evaluation metrics for Urdu text generation, like BLEU, METEOR, and ROUGE, primarily rely on n-gram overlaps, limiting their ability to capture semantic alignment beyond lexical matches. This is particularly evident in our DRS-to-text generation examples in Table 2. For instance, BLEU assigns a score of 16.67 to the generated translation ("*I am not ashamed yet*") compared to the gold reference ("*I'm not shy yet*") (Ex. 1, Table 2). While the generated text conveys the same core meaning, the BLEU score is low due to slight lexical variations in the choice of words like "ashamed" vs. "shy." This highlights BLEU's emphasis on lexical overlap over capturing the overall meaning of the sentence.

Similarly, for the translation ("*I bought two dozen pencils*") compared to ("*I bought 24 pencils*"), both sentences convey the same meaning but are penalized due to different representations i.e., "two dozen" vs. "24." This exemplifies the metric's failure to acknowledge acceptable paraphrases or equivalent expressions in the target language, further underscoring its limitations in multilingual contexts. In both cases, SMATCH indicates complete semantic alignment with a score of 1.0, highlighting the gap in traditional metrics' sensitivity to semantic fidelity. METEOR, which improves on BLEU by considering synonym matching and stemming, does provide higher scores for the same example (43.31 vs. BLEU's 49.12), but it is not immune to limitations. METEOR still struggles with capturing fine-grained semantic differences, as seen in Ex. 5 in Table 2, where the score of 37.04 fails to distinguish between "*He was born in 198X*" and "*He was born in the eighties*". Despite both sentences being semantically similar, METEOR's score is lower because it does not consider the subtleties of temporal expressions in Urdu and fails to fully match the corresponding time entities. The chrF score (which focuses on character-level n-gram overlap) in this context, with scores ranging from 21.95 (Ex. 1) to 39.79 (Ex. 2), similarly fails to capture the underlying semantic similarity. While chrF is more effective for languages with complex morphology, such as Urdu, it still penalizes minor differences in word structure and morphology, even when the generated text accurately conveys the intended meaning. In Ex. 1, "ashamed" vs. "shy" shows small morphological differences that affect chrF's performance, despite the generated text being semantically correct.

BERTScore, which attempts to measure seman-

tic similarity using pre-trained language models, is better suited for capturing the deeper semantic relationships between words. However, even this metric struggles when dealing with syntactic and morphological variations in Urdu. For instance, in Ex. 3, "*Everyone left*" and "*All have now left*" exhibit a difference in tense and aspect, yet the meaning remains intact. BERTScore performs better here with scores of 88.74, but still faces challenges when evaluating minor syntactic differences that do not affect the overall meaning.

These examples underscore the need for an evaluation approach that emphasizes semantic quality. Our GEN-PARS approach addresses this by focusing on whether generated texts preserve the semantic content of the original DRS. Across all examples analyzed, where traditional metrics like BLEU fluctuate significantly (ranging from 16.67 to 56.43), GEN-PARS achieves perfect SMATCH scores of 1.0 by parsing generated texts back to DRS. This validates semantic equivalence despite surface differences as evident in Table 6.

3 Methods and Results

This study presents two novel evaluation methodologies for assessing the quality of DRS parsing and generation in Urdu: (1) evaluating parsing through generation capabilities (PARS-GEN) and (2) assessing generation through semantic parsing (GEN-PARS). Unlike conventional metrics that often focus on surface-level text similarity or structural alignment, these methodologies offer a deeper, cross-task approach that assesses both structural and linguistic fidelity in Urdu semantic processing. To complement our cross-task evaluations, we also computed the Pearson correlation between metrics across the PARS/PARS-GEN and GEN/GEN-PARS evaluations. This correlation analysis helps us understand the relationship between structural accuracy (e.g., SMATCH F1 scores) and linguistic quality metrics (e.g., BLEU, METEOR, BERTScore).

PMB[2] is a multilingual dataset comprising semantic representations in English, Italian, German, Dutch, and Chinese. Leveraging the language-neutral nature of DRS, we transformed English DRS-Text pairs into Urdu through a systematic approach involving syntactic structure, concept and word alignment, grammatical genders, and cross-lingual adaptation through named entities. This methodology resulted in the first comprehensive semantic resource for Urdu, comprising 3,000 gold-standard (fully manually annotated) data instances. The dataset transformation employed a hybrid methodology: DRS transformations utilized rule-based techniques and human annotation, while text translations were generated using Google Translate API. The dataset was partitioned into 1,200 training, 900 development, and 900 test examples. To enhance dataset diversity and complexity, we applied multi-dimensional augmentation strategies, including named entities, lexical (encompassing common nouns, adjectives, adverbs, and verbs), and grammatical augmentations. This approach expanded the dataset to 10,800 training examples, supplemented by 6,857 silver (partially manually annotated) instances.

For bidirectional evaluation—converting PARS to PARS-GEN and vice versa—we employed byT5-based parsing and generation models, fine-tuned using our comprehensive augmented dataset[3]. We implemented a two-stage fine-tuning strategy consistent with (van Noord et al., 2020). The first stage involved fine-tuning the model on silver data for 3 epochs to establish foundational DRS knowledge. The second stage focused on gold data fine-tuning for 10 epochs. Experimental parameters included AdamW optimizer, polynomial learning rate decay ($1e{-}4$), batch size of 32, maximum sequence length of 512, and GeGLU activation function. These models achieved state-of-the-art performance in Urdu DRS processing, facilitating reversible data generation.

For the PARS/PARS-GEN evaluations in Urdu (see Table 3), we achieved a SMATCH F1 score of 79.77, indicating a moderate level of structural accuracy in parsing Urdu texts into DRS. When this parsed DRS output was subsequently evaluated through generation (PARS-GEN), performance varied across different metrics, highlighting the challenges posed by Urdu's morphological complexity. Notably, the PARS-GEN evaluation returned a BLEU score of 45.48, a METEOR score of 41.39, chrF of 40.57, BERTScore of 85.36, and ROUGE of 49.55. Among these metrics, BERTScore showed the highest correlation

[2]The PMB is developed at the University of Groningen as part of the NWO-VICI project "Lost in Translation—Found in Meaning" (Project number 277-89-003), led by Johan Bos. Urdu PMB is not part of the official website yet, but can be provided freely for scientific purposes.

[3]Our Urdu semantic parsing and text generation models are publically available for research purposes.

with the PARS structural evaluation (SMATCH), suggesting that it better captures semantic consistency across the tasks. However, lower scores in BLEU, METEOR, and chrF reflect the challenge of generating text that matches reference structures while accounting for Urdu's flexible syntax and morphology.

PARS	PARS-GEN				
S-F1	BLU	MET	chrF	B_Scr	RUG
79.77	45.48	41.39	40.57	85.36	49.55

Table 3: Experimental results of PARS and PARS-GEN on standard test sets for Urdu. Underlined are the results with hightest correlation. Note: S-F1 = SMATCH F1-Score; BLU = BLEU; MET = METEOR; B_Scr = BERTScore; RUG = ROUGE.

In the GEN/GEN-PARS evaluations (see Table 4), we assessed how well the generated Urdu text preserved the intended DRS semantics by parsing it back into a DRS representation. Here, the GEN approach achieved moderate scores, with BLEU at 53.31, METEOR at 53.07, chrF at 51.49, BERTScore at 88.33, and ROUGE at 59.40. The GEN-PARS evaluation returned a SMATCH score of 74.83, emphasizing that maintaining full semantic accuracy is challenging in text-to-DRS parsing for Urdu, possibly due to its unique syntactic structures. BERTScore again showed the strongest correlation with GEN-PARS results, indicating it is more aligned with the structural preservation needed in semantic evaluations.

GEN					GPAR
BLU	MET	chrF	B_Scr	RUG	S-F1
53.31	53.07	51.49	88.33	59.40	74.83

Table 4: Experimental results of GEN and GEN-PARS approaches on standard test sets for Urdu. Underlined are the results with hightest correlation. Note: GPAR = GEN-PARS; S-F1 = SMATCH F1-Score.

These results underscore that traditional metrics alone may not fully capture the linguistic intricacies in Urdu DRS parsing and generation. The relatively lower scores in some linguistic metrics, such as BLEU and METEOR, indicate that while structural preservation (PARS) aligns moderately well with these scores, morphological and syntactic differences specific to Urdu lead to lower alignment in n-gram-based and surface-level metrics. This suggests the potential benefit of incorporating additional language-specific evaluation strategies when working with morphologically complex languages like Urdu.

4 Analysis and Discussion

To further emphasize the usefulness of the reversible evaluation approaches, we have analyzed examples present in Table 1 and Table 2 by performing the reverse evaluations, i.e., PARS through PARS-GEN and GEN through GEN-PARS. Furthermore, we have performed Pearson correlation analysis on the reversible evaluation measures.

Reversible Evaluation Measures: While Sections 2 highlighted the limitations of traditional parsing and generation metrics individually, in this section we present the cases where our proposed evaluation approaches (PARS-GEN and GEN-PARS) provide complementary evidence of semantic and structural preservation. Through detailed analysis, we demonstrate how low scores in one type of evaluation (PARS or GEN) can be counter-verified by evaluating it in the reverse direction, revealing semantic equivalences that would have been missed.

Evaluating PARS through PARS-GEN: In analyzing DRS with structural overlap metrics like SMATCH, certain limitations in capturing the full semantic equivalence between the gold standard and generated DRS is evident. Table 1 highlights this issue through examples where PARS (DRS) scores do not adequately reflect semantic alignment despite the intended meaning being correctly represented. These examples underscore a critical drawback of relying solely on structural metrics, as they may fail to capture essential meaning alignment between generated and gold structures.

To address these limitations, PARS-GEN (text generation from DRS) evaluations in Table 5 supplement structural assessments with semantic overlap metrics, including BLEU, METEOR, ROUGE, chrF, and BERTScore, which provide a finer-grained view of how well the generated text aligns with the gold text. In Ex. 1, PARS-GEN achieves a BERTScore of 97.30 and METEOR of 69.14, capturing the semantic fidelity of the phrase "*Tom bought a new pickup*". Although SMATCH did not register structural similarity, the text-based evaluations in PARS-GEN reveal a strong overlap in meaning. Similarly, Ex. 2 achieves perfect PARS-GEN score across all metrics (BLEU:

Ex. No.	PARS DRS	PARS (SMATCH)	PARS GEN Text	Gold Text	GEN Scores				
					BLEU	METEOR	ROUGE	chrF	B_Scr.
1	male.n.02 Name “ٹام” new.a.01 AttributeOf +1 pick_up.n.01 buy.v.01 Agent -4 Beneficiary -3 Theme -1 Time +1 time.n.08 TPR now	00.00	ٹام نے ایک نیا پک اپ خریدا۔ (“Tom bought a new pickup.”)	ٹام نے ایک نیا پک اپ خریدا۔ (“Tom bought a new pickup.”)	71.43	69.14	71.43	64.14	97.30
2	male.n.02 Name “ٹام” female.n.02 Name “مریم” female.n.02 ANA -1 dog.n.01 Owner -1 photo.n.01 Creator -1 show.v.01 Agent -4 Recipient -3 Recipient -1 Time +1 time.n.08 TPR now	69.23	ٹام مریم کو اپنے کتے کی تصویر دکھاتا ہے۔ (“Tom shows Mary a picture of his dog”)	ٹام مریم کو اپنے کتے کی تصویر دکھاتا ہے۔ (“Tom shows Mary a picture of his dog.”)	100	99.93	99.99	100	100
3	person.n.01 EQU speaker neck.n.01 hurt.v.01 Patient -2 Patient -1 Time +1 time.n.08 EQU now	60.00	میری گردن میں اب بھی درد ہے۔ (“My neck still hurts.”)	آج میری گردن میں درد ہے۔ (“Today I have a pain in my neck.”)	57.14	61.47	61.54	48.07	87.11
4	quantity.n.01 EQU 30 person.n.01 Quantity -1 arrest.v.01 Patient -3 Time +1 time.n.08 TPR now	00.00	تیس افراد کو گرفتار کر لیا گیا۔ (“Thirty people were arrested.”)	تیرہ افراد کو گرفتار کر لیا گیا۔ (“Thirteen people were arrested.”)	56.43	54.35	61.54	61.72	94.55
5	person.n.01 EQU speaker have.v.01 Pivot -1 Theme +2 Time +3 quantity.n.01 EQU +1 quantity.n.01 EQU +1 money.n.01 Quantity + time.n.08 EQU now	57.89	میرے پاس بہت پیسہ ہے۔ (“I have a lot of money.”)	میرے پاس بہت پیسہ ہے۔ (“I have a lot of money.”)	83.33	80.66	83.33	76.08	97.63

Table 5: Evaluating PARS through PARS-GEN by taking examples from Table 1. Note: B_Scr. = BERTScore.

100, METEOR: 99.93, ROUGE: 99.99, chrF: 100, BERTScore: 100), demonstrating that, despite SMATCH's inability to capture semantic alignment, PARS-GEN accurately reflects the intended message that the `Owner` showed the `Recipient` a picture of dog.

Furthermore, Ex. 3 in Table 1 highlights a nuanced challenge where SMATCH (60.00) underestimates the semantic alignment due to complex relational and sentiment-bearing expressions. Here, the DRS encodes the phrase “*My neck still hurts*” yet this overlap is inadequately represented by the structural metric. In contrast, PARS-GEN scores in Table 5, with a BERTScore of 87.11, provides a closer approximation of the intended meaning, thereby validating the DRS from a semantic standpoint. Similarly, Ex. 5 also demonstrates this phenomenon, where a SMATCH score of 57.89 misses subtle lexical differences in phrases like (“*I have a lot of money*”), PARS-GEN BLEU (83.33) and BERTScore (97.63) confirm semantic equivalence, which structural evaluation alone failed to capture.

This analysis reveals that PARS-GEN complements structural metrics by providing a more robust measure of semantic fidelity in text generation tasks. By using both PARS and PARS-GEN, we gain a comprehensive understanding of meaning overlap, particularly in cases where linguistic nuances or variations may obscure the structural alignment but are nonetheless captured through text-based evaluations. Together, PARS and PARS-GEN offer a dual approach that effectively bridges the gap between structural and semantic overlap, enhancing the accuracy and reliability of DRS evaluation.

Evaluating GEN through GEN-PARS: The evaluation of generated text against gold DRS (after performing GEN-PARS) using semantic overlap metrics reveal critical insights into the limitations of traditional automatic metrics for text generation. Table 2 outlines these issues, showcasing several examples where semantic alignment is assessed through automatic word-overlap-based measures, e.g., BLEU, METEOR, ROUGE, chrF, and BERTScore. This discrepancy suggests that, traditional evaluation metrics for Urdu text focus on n-gram matching, they may not adequately capture the semantic richness and structural sequences represented in the DRS.

Transitioning to Table 6, which focuses on structural overlap metrics, we observe the implementation of GEN-PARS, which assesses the generated text against the original DRS. Notably, all examples (1-5) yield a perfect SMATCH score of 100, signifying that the generated structures align perfectly with the gold DRS. For instance, in Ex. 1, the transition from “*I'm not shy yet*” in GEN to the corresponding GEN-PARS representation maintains the event structure intact, reinforcing the idea that the generated text retains all necessary elements for a correct DRS encoding.

Furthermore, Ex. 3 and Ex. 4 reveal similar patterns. Both examples demonstrate that the generated text aligns seamlessly with the DRS structure, as evidenced by the SMATCH scores of 100. The transformation from “*All have now left*” and “*John gave the money to Mary*” to their DRS representations encapsulate the essential semantic components, reinforcing the effectiveness of GEN-PARS

Ex. No.	GEN Text	GEN-PARS (DRS)	Gold DRS	GPARS (SMATCH)
1	میں ابھی تک شرمندہ نہیں ہوں. ("I'm not shy yet.")	person.n.01 EQU speaker ashamed.a.01 Experiencer -1 Time +1 NEGATION <1 time.n.08 EQU now	person.n.01 EQU speaker ashamed.a.01 Experiencer -1 Time +1 NEGATION <1 time.n.08 EQU now	100
2	میں نے 24 پنسلیں خریدیں. ("I bought 24 pencils.")	person.n.01 EQU speaker quantity.n.01 EQU 24 pencil.n.01 Quantity -1 buy.v.01 Agent -3 Theme -1 Time +1 time.n.08 TPR now	person.n.01 EQU speaker quantity.n.01 EQU 24 pencil.n.01 Quantity -1 buy.v.01 Agent -3 Theme -1 Time +1 time.n.08 TPR now	100
3	اب سب چلے گئے ہیں. ("All have now left.")	NEGATION <1 person.n.01 NEGATION <1 leave.v.01 Theme -1 Time +1 time.n.08 TPR now	NEGATION <1 person.n.01 NEGATION <1 leave.v.01 Theme -1 Time +1 time.n.08 TPR now	100
4	جان نے رقم مریم کو دی. ("John gave the money to Mary.")	male.n.02 Name "جان" female.n.02 Name "مریم" money.n.01 give.v.03 Agent -3 Recipient -2 Theme -1 Time +1 time.n.08 TPR now	male.n.02 Name "جان" female.n.02 Name "مریم" money.n.01 give.v.03 Agent -3 Recipient -2 Theme -1 Time +1 time.n.08 TPR now	100
5	وہ امّی میں پیدا ہونے. ("He was born in 198X.")	male.n.02 time.n.08 YearOfCentury 198X bear.v.02 Patient -2 Time -1 Time +1 time.n.08 TPR now	male.n.02 time.n.08 YearOfCentury '198X' bear.v.02 Patient -2 Time -1 Time +1 time.n.08 TPR now	100

Table 6: Evaluating GEN through GEN-PARS by taking examples from Table 2. Note: GPARS = GEN-PARS

in maintaining structural integrity while providing a high-quality semantic output.

The role of word order and the presence of synonyms in model-generated outputs (either DRS or text) significantly influence the model performance and should be carefully considered. In the SMATCH evaluation, the impact of word order is generally minimal because SMATCH emphasizes structural overlap rather than the precise sequence of words. However, in cases where the meaning of a sentence is heavily based on its syntactic arrangement, SMATCH may not adequately capture the nuances, making it less effective for parsing evaluation. Similarly, SMATCH evaluates exact lexical entities, leading to penalties for synonymous expressions that maintain semantic equivalence but differ in lexical choice. To address these limitations, our cross-task evaluation approach (PARS/PARS-GEN) generates textual representations of DRSs and evaluates these using n-gram overlaps to assess word order and metrics like METEOR and BERTScore, which account for synonyms and contextual embeddings, respectively.

On the other hand, metrics such as BLEU, commonly used for evaluating text generation, impose strict penalties for variations in word order and the use of synonyms due to their reliance on n-gram-based overlap. To mitigate these issues, our counter-evaluation method for generation through parsing (GEN/GEN-PARS) transforms textual outputs into DRS representations, allowing evaluation through structural overlaps that are less sensitive to word order, as measured by SMATCH. This analysis elucidates the necessity of integrating both semantic (GEN) and structural (GEN-PARS) evaluations in understanding the quality of generated texts. While GEN metrics highlight the challenges posed by conventional evaluations in capturing semantic nuances, GEN-PARS effectively illustrates how generated structures can align with DRS, thus ensuring that the meaning is preserved. By leveraging both sets of metrics, we obtained a more nuanced view of the strengths and limitations of text-generation processes, fostering improvements in model training and evaluation methodologies.

Correlation Analysis: In evaluating DRS-based systems for Urdu, it is essential to analyze both quantitative performance measures and how well the system preserves underlying semantic content. Traditional metrics provide an initial foundation, but correlation analysis enables deeper insights into whether automatic evaluations effectively capture semantic quality and structural coherence. By analyzing correlations across automated measures such as PARS/PARS-GEN and GEN/GEN-PARS, we assess how reliably these metrics reflect true semantic accuracy in generated outputs.

We used Pearson correlation to examine the relationships between PARS/PARS-GEN and GEN/GEN-PARS scores. This analysis reveals the extent to which different metrics align—such as whether improvements in parsing accuracy correspond to enhancements in generation quality. A high positive Pearson correlation would indicate that the metrics consistently capture similar aspects of semantic and structural accuracy.

PARS/PARS-GEN Correlation: Our analysis for Urdu reveals statistically significant correlations across all metrics, despite the language's morphological complexity. BERTScore exhibited the highest correlation ($r = 0.2832$, $p < 4.55e\text{-}18$),

suggesting that neural-based metrics, like contextual embeddings, may more effectively capture semantic relationships in morphologically rich languages (see Table 7). This strong correlation with BERTScore implies that it could be particularly effective for evaluating the semantic quality of generated Urdu text, as it appears more sensitive to the subtle linguistic variations present in Urdu.

PARS vs. PARS-GEN	Corr-val	P-val
Pars vs BLEU	0.2318†	1.87e-12
Pars vs METEOR	0.1949†	3.69e-9
Pars vs ROUGE	0.2023†	9.12e-10
Pars vs chrF	0.2042†	6.25e-10
Pars vs BERTScore	<u>0.2832</u>†	4.55e-18

Table 7: Correlation results for PARS and PARS-GEN. Underlined values represent the strongest correlation. † indicates that the values are highly significant.

The remaining metrics also demonstrated significant, albeit weaker, correlations: BLEU (r = 0.2318), ROUGE (r = 0.2023), chrF (r = 0.2042), and METEOR (r = 0.1949). While these correlations are weaker, they remain highly significant, indicating that even traditional generation metrics can offer valuable insights into parsing performance. However, BERTScore's stronger correlation emphasizes the advantages of using contextual embeddings for capturing semantic fidelity in Urdu. The consistently positive and significant correlations across metrics affirm the reliability of our PARS-GEN approach for Urdu, demonstrating that parsing accuracy align well with generation quality metrics, with BERTScore emerging as particularly effective for assessing complex semantic content.

GEN/GEN-PARS Correlation: We extended the correlation analysis to GEN/GEN-PARS, examining how well generation metrics predict parsing performance, adding a complementary perspective on the relationship between these processes. BERTScore demonstrated the highest correlation in the GEN/GEN-PARS evaluation (r = 0.4073, p < 2.75e-37), indicating a moderate and highly significant relationship between generation quality and parsing accuracy (see Table 8). This high correlation suggests that neural-based embeddings are particularly effective at preserving semantic content that can be recognized by parsing models, even when dealing with morphologically rich languages. BLEU followed with a notable correlation (r = 0.3414, p < 5.36e-26), further highlighting its utility as a predictor of parsing performance.

GEN vs. GEN-PARS	Corr-val	P-val
BLEU vs Gen-Pars	0.3414‡	5.36e-26
METEOR vs Gen-Pars	0.2936‡	2.30e-19
ROUGE vs Gen-Pars	0.3043‡	9.82e-21
chrF vs Gen-Pars	0.2987‡	5.25e-20
BERTScore vs Gen-Pars	<u>0.4073</u>‡	2.75e-37

Table 8: Correlation results for GEN and GEN-PARS. Underlined values represent the strongest correlation. ‡ shows that the values are highly significant.

Other metrics also demonstrated significant correlations, albeit to a lesser extent. ROUGE (r = 0.3043), chrF (r = 0.2987), and METEOR (r = 0.2936) maintained positive and statistically significant correlations. These findings suggest that even traditional generation metrics capture some degree of semantic alignment in Urdu, but neural metrics like BERTScore remain more robust.

5 Conclusion

DRS parsing and generation are reversible processes that can be exploited in cross-task evaluations. Traditional metrics often fall short in capturing the true structural and linguistic quality required for accurate assessment. To address the limitations, we introduced two complementary methodologies, PARS-GEN and GEN-PARS, which offer a bidirectional framework to evaluate Urdu DRS processing more holistically. The PARS-GEN approach assesses parsing quality by generating text from parsed DRS, revealing linguistic nuances that purely structural metrics may miss. In parallel, GEN-PARS transforms generated text back into DRS, providing a structural and semantic evaluation of generation quality that goes beyond surface evaluations. Applying these methods to Urdu has yielded significant insights: (i) Urdu exhibits stronger correlations between generation quality and parsing accuracy than the reverse, indicating that high-quality generation is a reliable predictor of parsing performance; (ii) BERTScore shows the highest correlations, demonstrating their effectiveness in capturing Urdu's complex linguistic features; and (iii) The positive, statistically significant correlations across both evaluation directions validate the bidirectional parsing-generation relationship for Urdu.

Limitations The cross-task evaluations conducted for DRS parsing and generation offer a foundational approach to assessing the structural and linguistic quality of Urdu semantic processing comprehensively. However, the transformation process from DRS to text and text to DRS relies heavily on the capabilities of the underlying pre-trained language models. These models must demonstrate sufficient generalizability and robustness to achieve accurate and high-quality data transformations between DRS and text formats. Model biases or limitations in the pre-trained architecture may adversely impact performance, potentially resulting in evaluations that deviate from gold-standard outputs. This reliance on model quality underscores the need for continued refinement and bias mitigation in pre-trained models to ensure reliable and unbiased semantic transformation and evaluation.

Acknowledgments

The research is conducted at the Department of Computer Science, University of Turin, Italy, and is partially funded by the "HARMONIA" project (M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR). The project is supported under the NextGenerationEU program, with the funding identification details CUP C63C22000770006 - PE PE0000013. We extend our gratitude to prof. Viviana Patti, the principal investigator of the HARMONIA research initiative, for facilitating the funding of this work.

References

Muhammad Saad Amin, Luca Anselma, and Alessandro Mazzei. 2024. Exploring data augmentation in neural DRS-to-text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2178, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Saad Amin, Alessandro Mazzei, and Luca Anselma. 2022. Towards data augmentation for DRS-to-text generation. In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, November 30th, 2022*, volume 3287 of *CEUR Workshop Proceedings*, pages 141–152. CEUR-WS.org.

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.

Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2009. Urdu and the modular architecture of pargram. In *Proceedings of the Conference on Language and Technology*, volume 70.

Johan Bos. 2023. The sequence notation: Catching complex meanings in simple graphs. In *15th International Conference on Computational Semantics*, pages 195–208. Association for Computational Linguistics (ACL).

Miriam Butt and Tracy Holloway King. 2002. Urdu and the parallel grammar project. In *COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization*.

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.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Kasia M Jaszczolt and Katarzyna Jaszczolt. 2023. *Semantics, pragmatics, philosophy: a journey through meaning*. Cambridge University Press.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2010. Discourse representation theory. In *Handbook of Philosophical Logic: Volume 15*, pages 125–394. Springer.

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.

Rik van Noord. 2019. *Neural boxer at the IWCS shared task on DRS parsing*. in Proc. IWCS Shared Task on Semantic Parsing, Gothenburg, Sweden. Association for Computational Linguistics.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. 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.

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.

Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rik van Noord, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In *Proc. 13th International Conference on Computational Semantics-Short Papers*, pages 24–31. Association for Computational Linguistics (ACL).

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

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.