Indian Grammatical Tradition-Inspired Universal Semantic Representation Bank (USR Bank 1.0)

Soma Paul[1,*] **Sukhada Sukhada**[2,**] **Bidisha Bhattacharjee**[3,*] **Kumari Riya**[4,**]

Sashank Tatavolu[5,*] **Kamesh R**[***] **Isma Anwar**[6,*] **Pratibha Rani**[*]

[*]IIIT Hyderabad [**]IIT (BHU), Varanasi [***]SIST, Chennai

[1]soma@iiit.ac.in, [2]sukhada.hss@iitbhu.ac.in, [3]bidisha.bhattacharje@research.iiit.ac.in,

[4]kumaririya.jra.hss24@iitbhu.ac.in, [5,6]{sashank.tatavolu, isma.anwar}@research.iiit.ac.in

Abstract

In this paper, we introduce **USR Bank 1.0**, a multi-layered, text-level semantic representation framework designed to capture not only the predicate-argument structure of an utterance but also the speaker's communicative intent as expressed linguistically. Built on the Universal Semantic Grammar (USG), which is grounded in Pāṇinian grammar and the Indian Grammatical Tradition (IGT), USR systematically encodes semantic, morpho-syntactic, discourse, and pragmatic information across distinct layers. In the USR generation process, initial USRs are automatically generated using a dedicated USR-builder tool and subsequently validated via a web-based interface (SAVI), ensuring high inter-annotator agreement and semantic fidelity. Our evaluation on Hindi texts demonstrates robust dependency and discourse annotation consistency and strong semantic similarity in USR-to-text generation. By distributing semantic-pragmatic information across layers and capturing the speaker's perspective, USR provides a cognitively motivated, language-agnostic framework with promising applications in multilingual natural language processing.

1 Introduction

This paper introduces USR Bank 1.0, a multi-layered linguistic resource designed to capture not only the semantic content (predicate-argument structure meaning) of an utterance, but also the communicative intent of the speaker as it is expressed through linguistic expressions. While many existing semantic representation frameworks focus on abstracting away from surface-level grammar to model a deep, singular meaning, the novelty of USR is evident in the representation of the nuanced communicative intent of the speaker. Rooted in Indian Grammatical Tradition (IGT) (Sukhada and Paul, 2023; Garg et al., 2023) and Pāṇinian grammar (Sukhada et al., 2023), the Universal Semantic Representation (USR) framework aspires to closely maintain a systematic link to the surface structure through annotating vivakṣā — the speaker's perspective on what to express, how to express it, and to what extent.

The multi-layered design of USR is chosen not only to represent information of different linguistic strata, such as lexical, intra-sentential dependency relations and discourse level information, as is normally done in other Semantic Representation (SR) systems. The multi-layeredness in USR is uniquely a design need to distribute information bundled in one linguistic expression across different layers based on their semantic-pragmatic implication. For example, expressions like 'additionally' and 'along with' share the propositional meaning of the logical operator 'and' (conjunction). However, a speaker's selection of these more elaborate terms introduces a specific pragmatic implicature (e.g., 'inclusive' or 'additional'). The multi-layered structure of USR captures this distinction: the basic conjunction resides in one layer (Discourse), while the pragmatic import is explicitly isolated in another (Speaker's View). This decomposition makes the pragmatic-semantic contribution of an expression distinct yet interconnected within the holistic USR.

USR is a text-level representation that specifies disambiguated concepts along with their ontological categories and morpho-semantic information, such as plurality, tense, aspect, modality, and causativization, intra-sentential relations among these concepts through its syntactico-semantic annotation of kāraka relations, inter-sentential discourse relations and pragmatic information denoted by discourse particles, thus going beyond the semantics of predicate-argument structure used in traditional Semantic Representations.

We have successfully demonstrated natural language generation from USR for both Hindi and English, establishing a strong foundation for multilingual generation. Our ongoing efforts aim to extend

these generation capabilities to Tamil, Sanskrit, and other Indian languages. However, the present paper focuses exclusively on the creation and description of the Hindi USR Treebank. The strategic inclusion of Hindi and Sanskrit (Indo-Aryan), Tamil (Dravidian), and English (Germanic, within the larger Indo-European family) in our broader research program is intended to rigorously test the completeness, universality, and language-independent nature of the USR framework. These multilingual components, however, pertain to ongoing and future work and are not part of the dataset reported here.

In this paper, Section 2 introduces the Universal Semantic Grammar (USG) and its theoretical foundation in IGT. Section 3 elaborates on the multi-layered design principles of USR, including salient features that underscore its distinct contribution. Section 4 provides a concise review of existing Semantic Representations and their theoretical orientations, contextualizing USR's distinct contribution. Section 5 describes a comprehensive methodology employed for developing the USR Bank, detailing our semi-automatic annotation pipeline. Finally, Section 6 reports the Inter-Annotator Agreement (IAA) for USR annotation, along with an automatic evaluation of USR-to-text generation using automated and human annotators, offering empirical validation of the representation and annotation scheme's reliability.

2 USG: The Theoretical Framework of USR

The IGT framework conceptualizes language as an inherently holistic phenomenon. (Kiparsky, 2002) pointed out that Pāṇini's grammar organization is a device that starts from meaning information and incrementally builds up a complete interpreted sentence. In more concrete terms, the derivation of a sentence is initiated by constructing the morphosyntactic analysis, i.e., the arguments of a predicate (or events) are assigned syntactico-semantic roles (kārakas) based on the ontology of the events and the speaker's wish to express certain features of it (vivakṣā). Bhartṛhari (Iyer, 1965) compares language communication to painting: the speaker starts with a unified idea and expresses it part by part, with words interconnected by the principles of semantic compatibility (sāmarthya) to form a coherent whole. While existing semantic representation focuses on predicate-argument structure (who did what, where, etc.) (Abend and Rappaport, 2017), it does not capture the speaker's intention (vivakṣā), which shapes how events are expressed from the perspective of the speaker. For example, in the case of a simple event of "a boy's causing a glass to break", the conceptual structure based on the principle of semantic compatibility licenses an agent ("the boy") and a patient ("the glass") of the event 'breaking'. But, it depends on the speaker's communicative intent (vivakṣā) how he/she chooses to express this event. In the IGT framework, the kāraka roles, which determine the morpho-syntactic structure of the sentence, accordingly change. For example:

- In *"The boy broke the glass"*, the speaker foregrounds the agent, *"the boy"*, who functions as the kartā, the most independent participant of the event *break-0*.

- In contrast, in *"The glass broke"*, the speaker chooses to foreground the affected entity (*"the glass"*), thus making it kartā, the most independent participant of *break-1*.

The sub-eventive explanation of (Parsons, 1990) accounts for this analysis. Both break-0 and break-1 are sub-events of the larger event 'break'. Hindi uses two different lexical items for the two events: *toḍa* (break-0) and *ṭūṭa* (break-1). Thus, the semantic import of the dependency relations inspired by IGT and assigned to the arguments of predicates can conceptually be very different from what thematic roles or semantic roles of predicates convey (Kulkarni and Sharma, 2019).

3 Design of USR

Universal Semantic Representation (USR) is conceptualized as a multi-layered system designed for comprehensive meaning encoding. This system operates at three primary levels: a) lexico-conceptual - focusing on disambiguated concepts along with their semantic category; b) intra-sentential - detailing semantic relationships between head and dependents within a single sentence; and c) discourse - capturing inter-sentential coherence and anaphora (Garg et al., 2023). Additionally, USR incorporates an emerging pragmatic layer to capture linguistically expressed speaker's attitude or communicative intent.

The distinctive contribution of USR lies in the distribution of information between these layers: the lexico-conceptual layer establishes conceptual

<segment_id=(1-a)>

Concept	Index	Sem_cat	Morpho_sem	Dependency	Discourse	Speaker's view	CxN Comp.
Mohan	1	male				samāveśī[1]	2:begin
[ne_1]	2	per					15:op1
boy_1	3	anim/male				bhī_1[2]	15:op2
be_1-pres	4			0:main			
10	5	numex					7:count
inch	6						7:unit
[height_meas_1]	7			8:rmeas[3]			
tall_1	8		comparmore	4:kartā_samā-nādhikaraṇa			
$speaker	9	anim		10:genitive			
brother_1	10	anim/male		3:rv[4]			
$yad[5]	11			12:kartā			
come_1-past	12			12:rcdelim[6]			
Pune	13						14:begin
[ne_2]	14	place	10:source				
[conj_1]	15			4:kartā			

</segment_id>

[1]samāveśī – inclusive
[2]bhī_1 – also
[3]rmeas – **r**elation **meas**urement; measurement of event or entity
[4]rv – **r**elation ***vibhājana***; inequalities between two compared entities
[5]$yad – relative pronoun (all pronouns are prefixed by $)
[6]rcdelim – **r**elative **c**lause **delim**itation; when the relative clause delimits the head noun

Table 1: Representation of USR for (1-a).

anchors, the intra-sentential layer builds syntactico-semantic scaffolding over them, the discourse layer integrates these units into connected discourse, and the speaker's view overlays pragmatic intent.

To illustrate, consider the small discourse text given in Example (1) below. Table 1 and 2 present its USR.

(1) a. Along with Mohan, the boy who came from Pune is also 10 inches taller than my brother.
b. Besides that, they are also very strong.

The following sub-sections present the semantic-pragmatic analysis of this example text.

3.1 Lexico-Conceptual Layer

Every USR consists of a list of concepts: Simple or Complex Concepts (CC). Only entities, events, and modifiers, including quantifiers, are concepts. CCs represent higher-order cognitive schema that structure meaning independently of surface linguistic forms (Langacker, 1987; Evans and Green, 2018). For example, 10 inches (or 10") is [height_meas] CC in Table 1: Every simple concept is assigned a unique identifier (ID) that unambiguously specifies that concept. The digit with CC indicates the serial number of that CC in the USR. We can observe that the discourse particle 'also' is not represented as a concept in the concept column because it does not bear any propositional meaning. The relevant extra-propositional meaning (in this case 'inclusive') is added on "strong" in the Speaker's View column of Table 2. This implies that Mohan and the boy are "tall" as well as "strong".

This layer also includes ontological categories such as person, anim, place, season, day-of-week, week-of-month, month-of-year, male/female in the Sem_cat column (see Table 1) and records morpho-semantic information in Morpho_Sem (see Table 1) such as the comparative degree of an adjective (comparemore) on tall_1.

3.2 Intra-Sentential Layer

This layer encodes two kinds of information: (a) dependency relations among heads and dependents; (b) semantic tags for the components of Complex Concepts. The Dependency column of Table 1 and 2 illustrate the intra-sentential relations for (1-a) and (1-b), respectively.

According to IGT, there are two kinds of dependency relations: (a) kāraka relations, (b) kāraketara (other than kāraka) relations (Kulkarni and Sharma, 2019; Begum et al., 2008). *kāraka* roles include *kartā* (the most independent participant, often agentive), *karma* (the most desired object/patient), instrument, beneficiary, source and temporal-spatial.

<segment_id=(1-b)>

Concept	Index	Sem_cat	Morpho_sem	Dependency	Discourse	Speaker's View	CxN Comp.
$tyad	1				2:kartā	1.15:coref	
be_1-pres	2			0:main	1.4:conjunction	additional	
very_1	3			4:intf			
strong_1	4			2:kartā_samā-nādhikaraṇa		inclusive	

</segment_id>

Table 2: Representation of USR for (1-b).

There are 73 dependency relations in the current USR Guidelines V 4.2.1.

The main clause of (1-a) is a copulative sentence. Unlike most SRs that treat such predicative adjectives as a functor and the subject as its argument (e.g., tall(Mohan), tall(boy)), Pāniṇi's grammar treats the copula as the main predicate that indicates a state. That is why *be_1-pres* is assigned *0:main*. The noun that agrees with the copula is considered *expressed* (abhihita) and occupies the subject position, which is annotated as kartā in USR. The predicative adjective is annotated as kartā_samānādhikaraṇa[1]. This tag implies that the properties of 'boyhood' and 'tallness' reside in the same individual. Since in (1-a) both Mohan and the boy are tall, the kartā relation is specified on the CC ([conj_1]), which conjoins Mohan and the boy.

In addition, this layer specifies the internal composition of Complex Concepts. For example, in Table 1, Mohan and the boy are annotated as operands (op1, op2) of the CC [conj_1]. Similarly, the CC [height_meas_1] is internally structured into two components: count (10) and unit (inch), as indicated in the 'CxN Component' column. The next level of representation is the Discourse Layer.

3.3 Discourse Layer

In the discourse layer, we capture the semantics of discourse connectives. In (1-b), the author could have used the connective "and" in place of "besides that", which would have retained the discourse coherence of (1-a) and (1-b). However, the author has chosen the phrase "besides that" by which the author desires to express the conjunction and *something more*. In PDTB 3.0 Annotation Manual (Prasad et al., 2019), "besides" is annotated under Expansion. Conjunction, along with connectives such as "and" and "additionally." In contrast, USR differentiates between such connectives, recognizing that "besides" carries rhetorical weight beyond simple conjunction. Thus, the discourse layer highlights how the accumulation of meaning is shaped not only by propositional content but also by the speaker's rhetorical choices, which are further specified in the speaker's view layer.

3.4 Speaker's View

This layer, currently in its preliminary stage of development, aims to capture extra-propositional information overtly expressed through linguistic expressions in language. For example, in (1-a), the choice of "along with Mohan" instead of "Mohan and the boy" signals an inclusive nuance which is captured by the tag samāveśī 'inclusive' on Mohan in the Speaker's View column. In (1-b), the expression "besides that" specifies *adding to the list*. The tag 'additional' captures this meaning at the Speaker's View column (see Table 2) on the verb (1-b). In this way, the speaker's view layer complements the discourse layer, giving a fuller account of expressions like "besides that". The annotation scheme of this layer extends to other pragmatic categories, including definiteness (e.g., 'the' vs. 'a'), expressions of respect or formality, informal address, exclusive (e.g., only), and inclusive (e.g., also).

The present work examines how these nuanced pragmatic meanings are lexicalized and grammaticalized across languages, beginning with an initial comparative study of these categories in Hindi and English. This comparison reveals systematic and recurrent behavioral patterns—that is, regularities in how these pragmatic meanings are encoded, distributed, and triggered across constructions in the two languages. Such cross-linguistic regularities suggest that many of these pragmatic categories exhibit stable semantic–pragmatic behavior, making them strong candidates for universal modeling within USR.

The interplay between layers emerges as each layer contributes a distinct aspect of mean-

[1]The kartā_samānādhikaraṇa tag implies that kartā and its predicative adjective refer to the same entity.

ing—basic semantic content, discourse-level relations, and speaker-oriented nuances. Together, these layers create a holistic and robust representation, building meaning cumulatively from core concepts to complex relationships and speaker intent (For example see Figures 2 and 3 of Appendix A). Through this layered accumulation, USR achieves a rare balance between semantic abstraction and structural fidelity to natural linguistic expression.

4 Related Work

Most of the Semantic Representations (SRs) abstract away from surface-level grammatical and syntactic idiosyncrasies, focusing on the underlying meaning. A detailed overview and comparative analysis of various SR parameters can be found in (Boguslavsky, 2019). Some SRs are based on specific linguistic frameworks, which shape their representational choices and theoretical foundations. For example, Minimal Recursion Semantics (MRS) (Copestake et al., 2005) is based on Head-driven Phrase Structure Grammar (HPSG); the Prague Dependency Treebank (PDT) (Hajic et al., 2006) aligns with Functional Generative Description (FGD); FrameNet (Baker et al., 1998) is grounded in Frame Semantics; the Parallel Meaning Bank (PMB) (Abzianidze et al., 2017) in Discourse Representation Theory (DRT); and Abstract Meaning Representation (AMR) (Banarescu et al., 2013) adopts a neo-Davidsonian event-semantics. UMR (Universal Meaning Representation) (Van Gysel et al., 2021) extends PropBank (Palmer et al., 2005) and AMR into a unified framework that is designed to accommodate typologically diverse languages, including those with noun incorporation and affixal verb structures. It captures sentence-level predicate-argument structures, while also encoding features such as aspect, quantification, scope, pronouns, and multi-word expressions. At the document level, UMR models cross-sentential relations including co-reference, temporal ordering, and factuality.

4.1 USR and other Semantic Representations

With the above semantic representation systems (SRs) having existed for over a decade—and several still undergoing active development—the question naturally arises: why is there a need for yet another semantic representation system? We argue that the uniqueness of USR lies in two key aspects: 1.) the theoretical framework adopted for USR; and 2.) the distributive method of annotation of semantic–pragmatic information often bundled in one linguistic expression. By grounding the representation in Pāṇinian grammar and IGT, USR captures communicative intent (vivakṣā) and the layered interplay of concepts and propositions, enabling models to understand how a speaker intends to convey information in different contexts. This capability is crucial for generative and multilingual NLP systems, which rely on fine-grained semantic-pragmatic distinctions that existing SRs do not provide. The idea of decomposing the semantic-pragmatic meaning of an expression and representing it in different layers is unique to USR. In a recent work on PDT (Mikulová et al., 2025), the annotation schema of discourse particles in Czech is reported, where the pragmatic-semantic nature of these items is acknowledged. USR proposes a representation schema to capture this decomposition of meaning in a distributed manner.

5 Developing the USR Bank 1.0

This section describes the stages of the creation of USR Bank 1.0 and presents the statistics of USRs created so far.

5.1 Tool and Annotation

The development of USR Bank 1.0 follows a structured three-phase pipeline to ensure accuracy and efficiency.

5.1.1 Segmentation of Complex Sentences

As a pre-processing stage for the USR generation, a Segmentor is run on the input text that splits the text into sentences and further employs a principled segmentation strategy to handle complex or information-heavy sentences. Instead of treating the complex sentences as a whole, the Segmentor breaks them down into semantically coherent segments, typically each containing one finite clause. Segmentation adheres to consistent rules, such as splitting at discourse connectives, postulating elided elements, not segmenting relative clauses if the head noun is modified by one relative clause, and so on. Each segment is assigned a unique ID. Segment IDs accommodate *titles, headings*, and *fragments*, ensuring structural clarity throughout the annotation of a text. Evaluated against 500 gold-standard sentences, drawn from the NCERT Geography corpus, our Segmentor tool achieved an accuracy of 96.3%. An example of segmented

output is available in Table 3 for a sentence taken from the NCERT Geography textbook:

> <sent_id=Geo_11stnd_13ch_0039>
> Wave speed: It is the rate at which the wave moves through the water, and is measured in knots.

Sentence ID	Text
Geo_11stnd_13 ch_0039T	Wave speed
Geo_11stnd_13 ch_0039a	It is the rate at which the wave moves through the water.
Geo_11stnd_13 ch_0039b	And it is measured in knots.

Table 3: Segmented Output with appended specific segment ID

5.1.2 Automatic USR Generation using USR-builder

A USR-builder tool for Hindi has been developed to generate USRs automatically. The segments from the Segmentor tool are simultaneously fed into four NLP modules: (a) the Dependency Parser and Mapper that determines syntactico-semantic structures by identifying POS tags, head words and generating dependency relations between the head and its children; (b) Morphological analyzer that provides detailed morphological information such as root forms, tense-aspect-modality (TAM), gender, number, and person (c) the Named Entity Recognition (NER) tool that identifies and classifies named entities present in each segment; and (d) the Discourse Connective Marker Tool that operates on the whole input text to detect discourse connectives and establish relationships between different segments.

All linguistic information obtained from the aforementioned NLP tools is then fed to two concept identifier modules: (a) the Complex Concept Identifier tool and (b) the Simple Concept Identifier tool to identify atomic concepts and their associated grammatical features.

In the final stage, the outputs from all previous modules are passed to the Rule Applicator, which applies a predefined set of heuristics to integrate the linguistic and semantic information into the final USR format. The resulting USR captures the underlying semantics of the input text in a language-independent, human-readable and machine-interpretable format.

The Simple Concept Identifier, CC Identifier, and the Discourse Relation Marker tools have been developed in-house. The Complex Concept Identifier currently achieves an accuracy of 84.26%, while the Discourse Relation Marker demonstrates an accuracy of 94%. A schematic flowchart illustrating the overall architecture and data flow of the USR-builder is presented in Figure 1 in Appendix A.

5.1.3 Manual Validation via SAVI Interface

Once the USRs are automatically created, they are uploaded in the PostgreSQL database (Stonebraker et al., 1990). PostgreSQL is a powerful open-source relational database known for its robust support for complex queries, data integrity, and scalability. This makes it ideal for managing interconnected linguistic data and the semantic layers of USRs.

The database schema is hierarchical, linking a Chapter to Sentences, Sentences to Segments, and each Segment forming the base for Lexico-Conceptual, Construction, Relational, and Discourse tables. Manual validation of USRs is performed by trained annotators using the Semantic Annotation Validation Interface (SAVI), a custom-built, web-based interactive interface. SAVI significantly streamlines the validation process by adopting a multi-layered approach for organizing information into separate, intuitive tabs. This allows annotators to efficiently correct tags (e.g., Semantic_category, Morpho-Semantic, Speaker's View) via dropdown menus; validate dependency relations by selecting head indices and relation names, and confirm Complex Concept components (which are color-coded across tables for clarity). Furthermore, the features of the SAVI interface integrate visualizers for dependency trees and discourse graphs, providing immediate visual feedback that greatly aids in accurate validation.

5.2 Data

The existing dataset can be classified into parts. The first dataset is created and curated to understand various linguistic phenomena that need to be semantically represented in the USR. The second dataset evaluates how well the framework and representation work for real-world texts from specific domains. The current statistics for the annotated data in USR Bank 1.0 are given in Table 4, and the statistics of the top 5 most frequently annotated dependency relations are given in Table 5.

Count of	First Data	Health Domain	Education
Sentences	659	168	5727
Segments	659	261	7029
Simple Concepts	2809	2131	56734
Complex Concepts	356	437	6888

Table 4: USR Bank data statistics.

Dependency Relation	Frequency
Modifier (mod)	7579
Genitive relation (r6)	6888
kartā (k1)	6655
karma (k2)	3031
Location (k7p)	2563

Table 5: Top 5 most frequent Dependency Relations annotated in USR Bank.

5.2.1 First Data: Manually Curated Simple Sentences

The primary corpus for USR Bank 1.0 comprises 659 simple and small sentences. This data is created manually, with the focus on encoding information at various linguistic levels. The primary goal of this dataset is to provide a controlled environment for detailed linguistic annotation. Table 5 shows the statistics of the top 5 most frequent dependency relations annotated in the data.

5.2.2 Second Data: Domain-specific text (Health and education)

The second data set is taken from two different domains, namely health and education. The health data is derived from consent forms used for patients and their relatives undergoing specific medical procedures by Christian Medical College, Vellore. The data for the education domain is sourced from the NCERT (National Council of Educational Research and Training) and NIOS (National Institute of Open Schooling) geography textbooks in Hindi, ranging from Class 6 to 12. This dataset offers domain-specific, thematically coherent material, ideal for evaluating the adaptability and depth of the USR framework across real-world contexts.

5.2.3 Annotated Data Statistics

The current statistics for the annotated data in USR Bank 1.0 are given in Table 4, and the statistics of the top 5 most frequently annotated dependency relations are given in Table 5.

6 Evaluation

The USR Bank 1.0 is evaluated in this paper using two parameters: (i) ease of annotation and consistency in the annotation schema, and (ii) effectiveness of USR for a downstream application, namely, natural language text generation. For the first, we calculated the Inter-Annotator Agreement (IAA) score and reported it in Section 6.1. For the second, we evaluated the semantic fidelity of USRs by comparing the quality of texts generated from USRs - both manually by human annotators and automatically by a large language model - with the original source text. The underlying assumption is that the closeness of the text generated from the USR with that of the original text will prove the correctness of the meaning representation in USR. In this paper, all evaluations are done for Hindi.

6.1 Evaluation Parameter 1: Inter-Annotator Agreement (IAA)

To obtain a more fine-grained picture of consistency, we designed two IAA settings: the first focusing only on dependency and discourse layers to capture core structural agreement (refer to Table 6), and the second including all four layers (lexico-conceptual, dependency, discourse, and speaker's view) to evaluate the full complexity of USR annotation (refer to Table 7). The experiment was conducted on two carefully selected datasets comprising 70 (Table 6) and 105 (Table 7) unique segments, respectively. Each segment, averaging 11–12 words in length, was extracted from the NIOS and NCERT geography textbook corpora, after preprocessing with our Segmentor Tool. The USR Builder generated the initial USRs for these segments, which were then uploaded to the database for independent validation by human annotators.

Two groups of annotators were involved: the first group consisted of two experienced annotators who had been working with this representation scheme for over a year, while the second group comprised two relatively new annotators with about two months of experience. Each annotator independently worked on their assigned set without prior consultation. After completion, the annotations were systematically compared to measure inter-annotator consistency.

6.1.1 Result

Inter-Annotator Consistency was quantitatively measured using both raw agreement percentage and Cohen's Kappa (k). Cohen's Kappa provides a more robust measure of agreement by adjusting for the proportion of agreement that would be expected by chance. For composite annotations (like dependency relations, which involve both a head-dependent pair and a specific label), Cohen's Kappa is calculated by considering each possible combination of head, dependent, and relation label as an annotation unit, allowing for a standard application of the formula.

Features	Cohen's Kappa	Agreement %
Dependency	0.8465	0.8912
Discourse	0.8817	0.9978

Table 6: IAA results using Cohen's Kappa (k) and Agreement Percentage.

Features	Cohen's Kappa	Agreement %
Dependency	0.8020	82.63
Discourse	0.6030	89.81
Speaker's View	0.7164	90.48
Sem_Cat	0.8949	97.90
Morpho_sem	0.6861	92.50
Construction	0.7520	86.22

Table 7: IAA results using Cohen's Kappa (k) and Agreement Percentage for mentioned features.

Metric	A1	A2	A3	Gemini Model
Cosine-Similarity	0.8866	0.8277	0.8065	0.9347

Table 8: Semantic Similarity scores for annotators: A1, A2, A3, and Gemini Model.

The Inter-Annotator Agreement (IAA) analysis reveals the following patterns of mismatch in annotation across the two annotators. Variability was particularly evident in co-reference resolution, where one annotator consistently linked entities to their initial mention, while the other preferred the most proximate mention within the discourse. Similar variation was found in head selection for dependency relations such as taking the verbalizer of a complex predicate as the head while the construction label [cp] is to be taken as the head. In addition, there are differences in the way constructions have been identified. There are some instances involved the omission of semantic category labels and morpho-semantic relations, further contributing to annotation inconsistency.

Despite these issues, the results, summarized above, demonstrate a remarkably high level of consistency between the annotators for both dependency-level and discourse-level annotations. This strong agreement empirically affirms the clarity, unambiguous nature, and semantic groundedness of the USR guidelines and its tagset.

6.2 Evaluation Parameter 2: USR-to-Text Generation

The objective of this experiment was to evaluate the completeness and faithfulness of information represented in the USR by generating texts manually and automatically from a set of gold USRs.

For this experiment, we used a manually validated gold set of USRs from a Hindi medical consent form from the health domain containing 59 segments. We conducted experiments of manual generation and automatic generation. We report here the cosine similarity measure of each generation against the original text. Three annotators participated in the manual USR-to-text generation task, each independently producing texts from the same set of USR. These three annotators were new annotators who were trained in USR annotation for only one month at the time of the experiment. Automatic text generation was done using the Gemini 2.5 pro model (Gemini Team, 2023).

6.2.1 Result

We have used the multilingual sentence transformer model (paraphrase-multilingual-MiniLM-L12-v2) to evaluate the quality of the texts generated by the three annotators as well as by the Gemini model through the Cosine similarity measure. These results summarized in Table 8 demonstrate strong agreement between the texts generated by the annotators and the original text, with all three annotators achieving high similarity scores above 80%. Also, the above 90% similarity score shows very high similarity between the original text and the model-generated output.

The overall mean similarity scores across annotators indicate high semantic consistency in the annotated USRs. Inter-Annotator Agreement was simi-

larly robust, with pairwise similarities consistently above 80%, showing that all three annotators maintained comparable semantic fidelity to the source text while producing linguistically diverse alternatives. The higher similarity score for the model-generated output, however, can be attributed to its reliance on surface-level word matching, whereas human annotators focus on capturing the finer semantic and pragmatic nuances of the USR, often rephrasing or restructuring the text in ways that reduce lexical overlap with the original. These results suggest that the annotation protocol effectively captured the meaning of the original texts. Given that medical consent forms demand high precision and clarity for patient comprehension, this analysis demonstrates how well our USR-based generation approaches preserve semantic meaning, structural integrity, and adherence to the expected patterns of critical medical information. We are also investigating in more detail why the model-generated output achieves a higher similarity score than the human annotation.

7 Conclusion

The USR Bank 1.0 advances the field of semantic representation by systematically integrating key principles from the Indian Grammatical Tradition. Anchored in the Universal Semantic Grammar (USG) framework, it captures core concepts from IGT — namely, sāmarthya (semantic compatibility) and vivakṣā (speaker's intention) — to offer a multi-layered, coherent, and cognitively grounded model of textual meaning representation. Evaluations through inter-annotator agreement and USR-to-text generation have demonstrated the reliability and semantic consistency of the framework. Its successful application in Hindi and ongoing efforts to extend it to Tamil, Sanskrit, and English demonstrate its potential for cross-linguistic and multilingual generation. This work bridges classical linguistic theory with modern language technology, offering a scalable, language-agnostic semantic model. Future developments will focus on expanding the treebank across more languages and refining automatic USR construction tools to enhance multilingual NLP capabilities.

Limitations

The annotators require a good amount of training in Universal Semantic Grammar before starting the annotation. Retaining good annotators is an expensive affair.

Acknowledgment

We are thankful to every member of the Language Communicator Tool for End Users team, along with Kirti, Fatema, Arjun, Sudarshan, Hymavathi, Mohan, Rajni, Sabharaj, Satyaprakash, Shweta, Varshith, Manash, Manu, Sakshi, Muskan, Sanchari, Saumini, Vandana, and interns from Banasthali Vidyapeeth for their contributions to data preparation and experiments. We are grateful to MeitY, GoI, for supporting and funding the project.

References

Omri Abend and Ari Rappaport. 2017. The universal anaphora annotation framework. *Transactions of the Association for Computational Linguistics*, 5:561–577.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik Van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *arXiv preprint arXiv:1702.03964*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Igor Boguslavsky. 2019. Interlingual lexical ontology: Design, construction, applications. *Cham: Springer*.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3:281–332.

Vyvyan Evans and Melanie Green. 2018. *Cognitive linguistics: An introduction*. Routledge.

Kirti Garg, Soma Paul, Sukhada Sukhada, Fatema Bawahir, and Riya Kumari. 2023. Evaluation of universal semantic representation (USR). In *Proceedings of the Fourth International Workshop on*

Designing Meaning Representations, pages 13–22, Nancy, France. Association for Computational Linguistics.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Jan Hajic, Jarmila Panevová, Eva Hajicová, Petr Sgall, Petr Pajas, Jan Štepánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtskỳ, Magda Ševcíková-Razímová, and 1 others. 2006. Prague dependency treebank 2.0. *CD-ROM, linguistic data consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

K. A. Subramania Iyer. 1965. *The Vākyapadīya of Bhartṛhari (Kāṇḍa II)*. Motilal Banarsidass.

Paul Kiparsky. 2002. On the architecture of pānini's grammar. Three lectures delivered at the Hyderabad Conference on the Architecture of Grammar, Jan. 2002, and at UCLA, March 2002.

Amba Kulkarni and Dipti Misra Sharma. 2019. Pāṇinian syntactico-semantic relation labels. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 198–208.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*, volume 1. Stanford university press.

Marie Mikulová, Barbora Štěpánková, and Jan Štěpánek. 2025. From form to meaning: The case of particles within the prague dependency treebank annotation scheme. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2163–2175.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2019. The penn discourse treebank 3.0. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Michael Stonebraker, Lawrence A. Rowe, Michael Ling, and Jeffery W. Du Bois. 1990. The design of postgres. *ACM Transactions on Database Systems (TODS)*, 15(2):174–188.

Sukhada Sukhada, Sirisipalli Veera Hymavathi, and Soma Paul. 2023. Generation of mrs abstract predicates from paninian usr. In *Proceedings of the 30th International Conference on Head-Driven Phrase Structure Grammar, University of Massachusetts Amherst*, pages 122–142, Frankfurt/Main. University Library.

Sukhada Sukhada and Soma Paul. 2023. Theory of sāmarthya in Indian Grammatical Tradition: The foundation of Universal Semantic Representation. In *Int J Sanskrit Res*, pages 17–22.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, and 1 others. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

A Additional Content

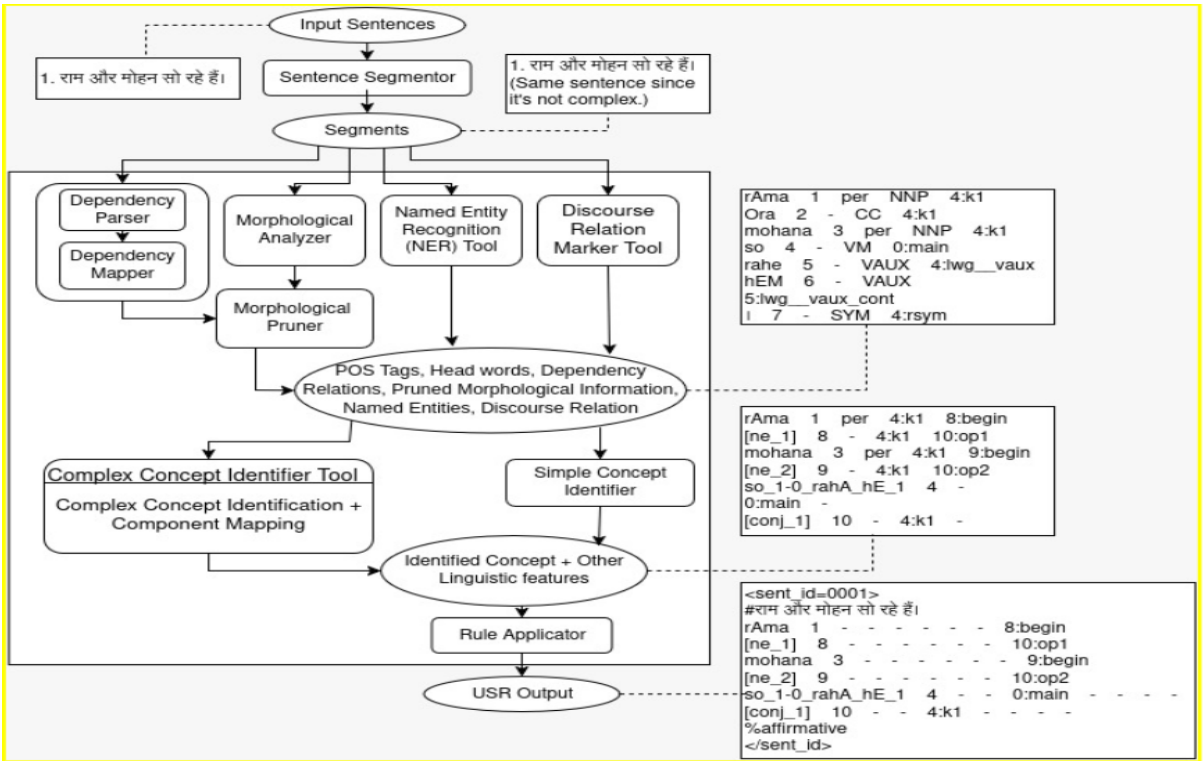

Figure 1: Architecture of USR Builder.

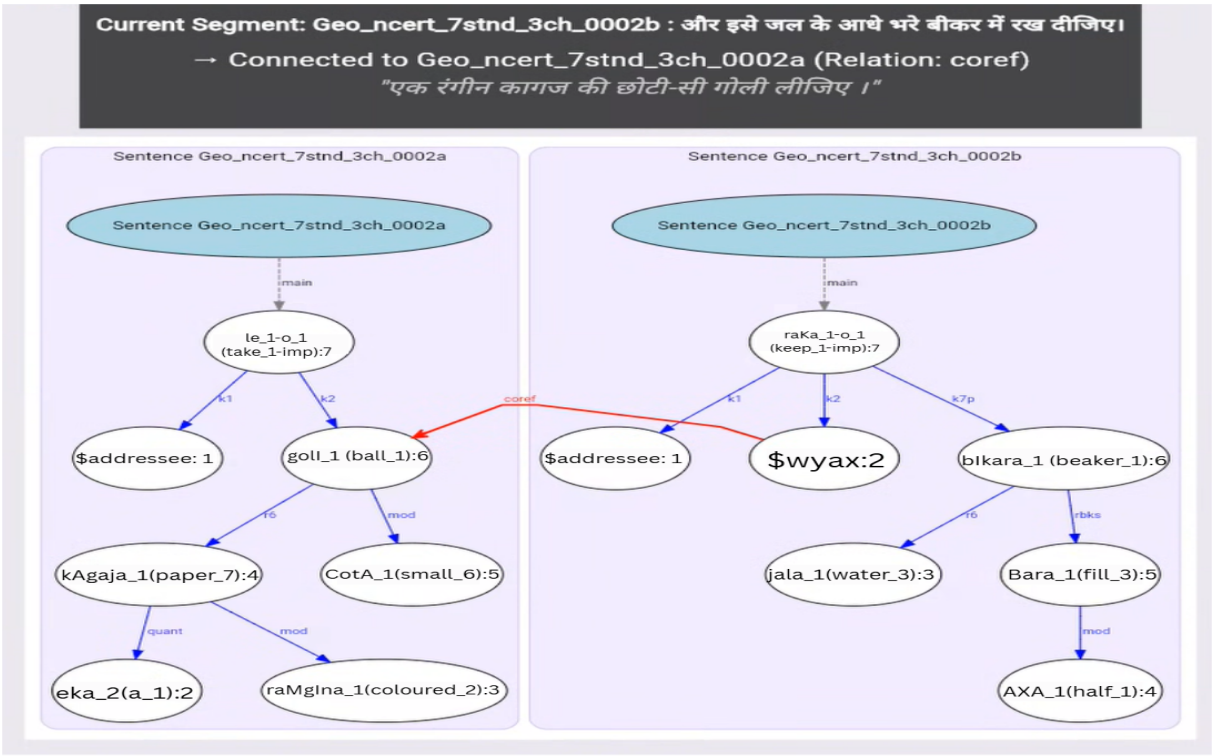

Figure 2: Annotation of inter segment co-reference.

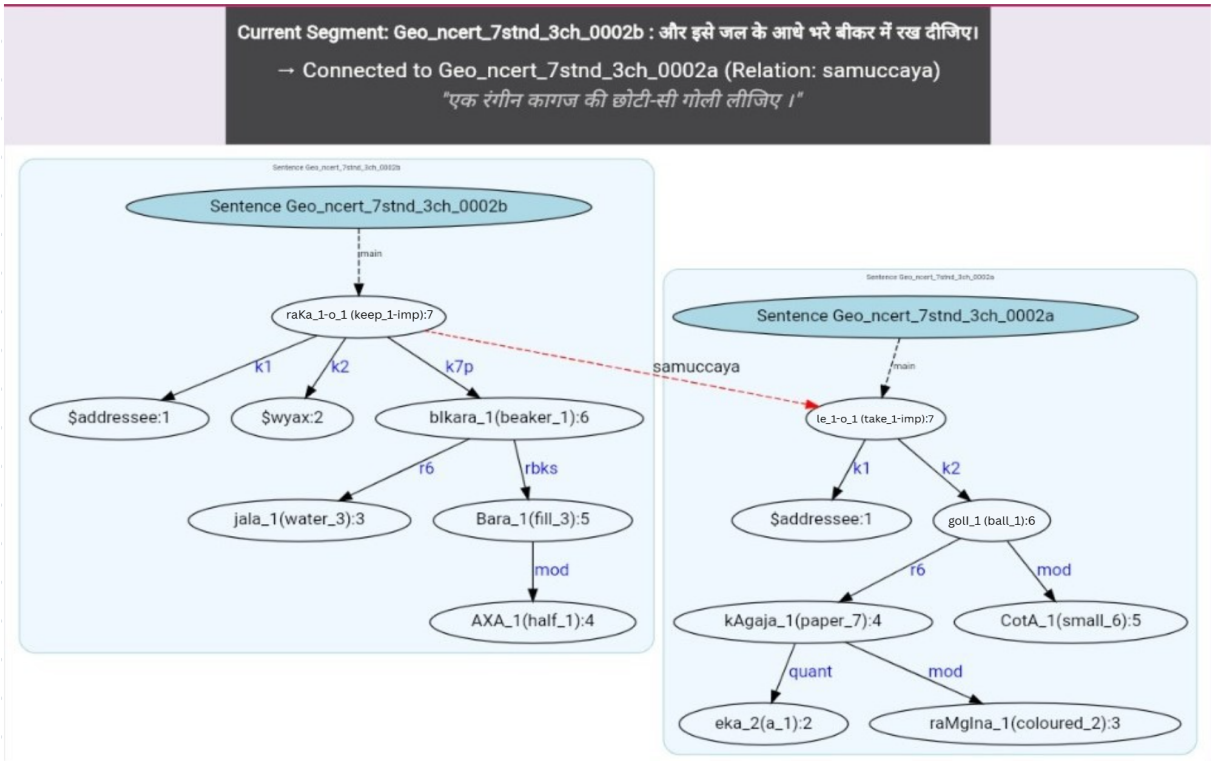

Figure 3: Annotation of dependency and discourse connective relations.