ISSN 2415-8712 (on-line); ISSN 1682-671X (print)
Shodoznavstvo, 2026, No. 97, pp. 183–228
doi: https://doi.org/10.15407/skhodoznavstvo2026.97.183

UDC 81’367.622:81’37:004.912

COMPOUND FORMATION AND REPRESENTATION FOR COMPUTATIONAL PROCESSING: A UNIVERSAL SEMANTIC REPRESENTATION APPROACH WITH A *PĀṆINIAN* PERSPECTIVE

Sudarshan Gautam
Research Scholar
Department of Humanistic Studies
Indian Institute of Technology (IIT) BHU,
Varanasi, 221005, India
sudarshangautam.rs.hss22@itbhu.ac.in
ORCID: 0000-0002-0707-3603

Sukhada
Assistant Professor
Department of Humanistic Studies
Indian Institute of Technology (IIT) BHU,
Varanasi, 221005, India
sukhada.hss@itbhu.ac.in
ORCID: 0000-0003-2224-0323

Understanding the semantic structure of compound formations is a central concern in both theoretical and computational linguistics. The Universal Semantic Representation (USR) provides a computationally tractable framework for capturing meaning in a structured, language-independent manner, drawing inspiration from the Indian Grammatical Tradition (IGT). This study applies Pāṇinian principles of compounding to the USR framework in order to analyse compound structures from a Pāṇinian perspective. While principles such as *ākāṅkṣā* (expectancy) and *yogyatā* (semantic congruity)

are foundational to semantic interpretation, the study particularly demonstrates how specific Pāṇinian constraints, including "*saviśeṣaṇānāṃ vṛttir na*" and "*vṛttasya vā viśeṣaṇayogo na*" help resolve modifier-scope ambiguities in automated systems and semantic parsing. The paper further examines the internal semantic organisation of compound expressions and the hierarchical relations that emerge in nested and complex compound constructions. Incorporating these insights into the USR framework demonstrates how compound semantics, dependency relations, and contextual interpretation can be systematically represented across languages for computational processing. Examples drawn from Hindi geography textbooks illustrate the practical applicability of the proposed analysis and underscore the relevance of Pāṇinian grammar for advancing computational approaches to semantic representation and Natural Language Generation (NLG). The study analyses 1,504 sentences from Hindi geography textbooks and observes that compound density reaches approximately 52 % in technical chapters, thereby necessitating a systematic logic for representing nested compound structures and preserving semantic integrity. The findings suggest that the integration of Pāṇinian grammatical insights with USR contributes significantly to Natural Language Processing (NLP) models for Indian languages by improving semantic precision, interpretive consistency, and computational representation of complex linguistic structures. The study highlights the continuing relevance of traditional Indian grammatical theories in contemporary computational linguistics and knowledge representation research.

Keywords: *vṛtti*, *Pāṇinian* grammar, *ekārthībhāva-sāmarthya*, compound formation, semantic compatibility, Universal Semantic Representation

1. Introduction

Language serves as the primary tool for expressing and structuring human thought and social behaviour. In this study, the term *language* is used in the restricted sense of human language, understood as a rule-governed symbolic system with specific cognitive and neurobiological foundations. Human language may be realised through different modalities, such as spoken, written, or signed forms, while remaining subject to the same underlying grammatical and semantic principles. Through language, humans articulate their internal thoughts and emotions, and meaning is systematically conveyed through sentences. Classical Sanskrit tradition does not offer a single, uniform definition of the sentence (*vākya*). Rather, multiple characterisations are found in the tradition, such as Kātyāyana's *vārtika* "*eka-tiṅ vākyam*" (a sentence consists of a single finite verb) [Patañjali, Kaiyaṭa, Bhaṭṭa

2018; Varadarājācārya 2004, *64*], "*sup-tiṅ-anta-cayo vākyam*" (a sentence is a collection of nominal and verbal inflected forms) [Amarasiṃha 1971], and "*pada-samūho vākyam*" (a sentence is a grouping of words) [Jha 2010, *94*; Ācārya 2014]. These definitions reflect different analytical perspectives, verbal, morphological, and structural, adopted by various scholars within the tradition. Although Pāṇinian[1] Grammar focuses on Sanskrit, it provides insights into global linguistic principles and computational models. Frits Staal argued that Pāṇini's work prefigures the concept of generative grammar introduced by Noam Chomsky in the 20th century. Pāṇini's rules can generate all the valid forms of Sanskrit words and sentences, similar to how generative grammar describes how sentences are constructed in natural languages [Staal 1972]. In everyday communication, humans comprehend and express meaning through sentences. The complete meaning of a sentence arises from the combination of the meanings of each component element used in the sentence. The Indian grammatical tradition has significantly contributed to the field of verbal cognition (*śābda-bodha)*. For *śābda-bodha*, the *anvaya* (connection) among the components of a sentence according to their *ākāṅkṣā* (expectancy), *yogyatā* (meaning congruity), *sannidhi* (contiguity), and *tāt-parya* (intention) is essential, as mentioned by Nāgeśa Bhaṭṭa in his *Paramalaghu-mañjūṣā:* "*atha śābda-bodha-sahakāra-kāraṇāni ākāṅkṣā-yogyatā-āsatti-tāt-paryāṇ*" [Bhaṭṭa 2006, *47*]. For instance, let us look at sentence (Ex. 1):

(Ex. 1) Sanskrit: *rāmo gṛhaṃ gacchati*
Gloss: Rama *nom. m. sg.* home *acc.-n. sg.* go *pres. 3 sg.*
English: Rama goes home.

This sentence has three elements: *rāma,* home (*gṛha)*, and √*gam*. Each of these three has its meaning: Rama, home, and motion. When all these elements are interconnected, their combination yields the meaning 'Rama goes home'. Let us first define what is meant by *ākāṅkṣā, yogyatā*, *sannidhi*, and *tātparya.*

[1] In this paper, 'Pāṇinian grammar' is understood in the broader traditional sense, encompassing the *muni-traya* (Pāṇini (the *sūtra-kāra*), Kātyāyana (the *vārttika-kāra*), and Patañjali (the *bhāṣya-kāra*)), as well as the subsequent grammatical tradition that follows them, and not Pāṇini alone as the author of the Aṣṭādhyāyī.

1.1. Ākāṅkṣā (Expectancy)

Ākāṅkṣā refers to a situation where a word cannot convey a complete sense on its own and requires another word. Nāgeśa Bhaṭṭa defined the *ākāṅkṣā* in the following manner:

"*vākya-samaya-grāhikā ākāṅkṣā sā caika-padārtha-jñāne tad-arthānvaya- yogyārthasya yaj jñānaṃ tad-viṣayecchā*" [Bhaṭṭa 2006, *47*].

"The listeners desire to know the other words or their meanings to understand the full message of the sentence".

A word has *ākāṅkṣā* for another if it cannot produce complete knowledge in a sentence without it. In every language, certain lexical items require other elements in order to yield a syntactically well-formed and semantically complete sentence. In this sense, "meaning completion" is understood as the formation of a sentence in formally definable terms. From the perspective of formal semantics, such semantic completeness corresponds to the formation of a proposition that can be evaluated as true or false. For example, a noun marked with the nominative case does not require a verb per se, but rather a predicate to complete its meaning; such predicates may be realised not only as verbs but also as adjectives or nouns, as in nominal sentences (e.g., *vīro rāmaḥ* 'Rāma is a brave'). Similarly, a verb such as *buy* conventionally selects both an external argument (the buyer) and an internal argument (the item purchased) in order to express a complete propositional meaning. A string of words such as *cow, bird*, *man* does not form a sentence and therefore fails to denote a truth value in the sense assumed in formal semantic theories. Because there is no connection between them due to a lack of fulfilment of *ākāṅkṣā*.

1.2. Yogyatā (Meaning Congruity)

Nāgeśa Bhaṭṭa, in his *Parama-laghu-mañjūṣā*, defines *yogyatā* as:

"*paras-parānvaya-prayojaka-dharma-vatvam yogyatvam*" [Bhaṭṭa 2006, *49*].

"*Yogyatā* is the quality or ability of a *pada* (inflected word) to communicate its meaning *(padārtha)* in a way that is logically related to other words".

The term *yogyatā* refers to the logical consistency of words within a sentence, ensuring they fit together to convey a meaningful message. It is about determining whether a sentence makes sense based on our understanding and experience. For example, in the sentence *jalena siñcati* (He sprinkles it with water), there is *yogyatā* because

wetting typically involves using a liquid like water, which is logical and makes sense. On the other hand, in the sentence *cagninā siñcati* (He sprinkles it with fire), there is no *yogyatā* because wetting with fire is illogical and doesn't match our understanding of how things work.

1.3. Sannidhi/Āsatti (Contiguity)

Sannidhi, or *āsatti*, essentially states that words in a sentence must be spoken immediately, without significant gaps. This ensures their uninterrupted connection and coherent apprehension. If words are spoken with long intervals or are interrupted by unrelated words, even *ākāṅkṣā* and *yogyatā* cannot establish their meaningful interrelation. For instance, if one word is uttered now and another word is uttered 10 minutes later to form a sentence, the lack of continuity *(vyavadhāna)* between the two words would prevent coherence, making it impossible to derive any meaningful interpretation *(sārthaka-vākyārtha)* from them. Nāgeśa Bhaṭṭa defines it as follows:

"*prakṛtabodha-anukūla-pada-avyavadhānam āsattiḥ*" [Bhaṭṭa 2006, *49*].

"*Āsatti* is the uninterrupted connection between words that facilitates the intended relational understanding (*anvaya-bodha*)".

This means that *āsatti* refers to the absence of interruptions caused by intervening words that are opposed to, or irrelevant to, the intended relational understanding (*anvaya-bodha*). For example, look at (Ex. 2).

(Ex. 2) Sanskrit: *[2]*giriḥ bhuktam agnimān devadattena*

Gloss: mountain *nom. Sg* eaten *acc. m/n. sg* fiery *nom. sg*. Devadatta *ins. m. sg.*

English: *The mountain, is eaten, fiery, by Devadatta.

In (Ex. 2), the words *giriḥ* (mountain) and *agnimān* (fiery) stand in a *viśeṣaṇa-viśeṣya-bhāva* (modifier-modified relationship). In the tradition of Nāgeśa Bhaṭṭa, such relations are said to be optimally interpreted when the relevant elements are contiguous, as in *giriḥ agnimān* (fiery mountain), *bhuktam devadattena* (is eaten by Devadatta). From this perspective, contiguity enhances clarity in relational understanding (*anvaya-bodha*). It should be noted, however, that this is not a

[2] The asterisk (*) preceding a sentence indicates that the expression is considered grammatically invalid or unacceptable in linguistic analysis.

general constraint on natural languages: modifier-modified relations may also be established across intervening material, as extensively documented in the literature on long-distance dependencies. Sanskrit, including Vedic Sanskrit, likewise exhibits such non-contiguous dependencies.

But in (Ex. 2), the word *bhuktam* (eaten) disrupts the connection between *giriḥ* (mountain) and *agnimān* (fiery). Thus, the lack of *āsatti* makes it harder to achieve immediate and precise verbal cognition (*śābda-bodha*). Nāgeśa Bhaṭṭa holds a slightly different view regarding the concept of *āsatti* (contiguity). As said, "*āsattir api manda-buddheḥ avilambena śābda-bodhe kāraṇam, amanda-buddhes tu āsatti-abhāve'pi padārthopasthitau ākāṅkṣāditaḥ avilambenaiva bodho bhavati*". That means, "*āsatti* (contiguity) is helpful for individuals with slower intellect (*manda-buddhi*) and for a person with ordinary intelligence to understand the meaning of a sentence (*vākyārtha*)" [Bhaṭṭa 2006]. However, those who are intellectually advanced (*vyutpanna*) or analytical can grasp the meaning of a sentence even in the absence of *āsatti*, despite interruptions by unrelated words. They rely solely on their understanding of the elements (*padārtha-jñāna*) and the strength of contextual expectations (*ākāṅkṣā*) to derive the meaning of the sentence. Therefore, *āsatti* does not hold as much significance for such intellectually capable individuals.

1.4. Tātparya (Intention of the Speaker)

Nāgeśa Bhaṭṭa defined *tātparya* in the following manner:

"*Etad vākyaṃ padaṃ vā etad-artha-bodhāyoccāraṇīyam iti īśvarīyecchā tātparyam*" [Bhaṭṭa 2006, *52*].

" 'The sentence or phrase should be pronounced to communicate its meaning'. This type of divine wish is called *tātparya*".

In common parlance, regarding words with multiple meanings, 'I pronounce this word or sentence to convey this meaning', such a type of speaker's intention is called (*tātparya*). In popular understanding, knowledge of context and situation is essential for determining the speaker's intent. Therefore, when someone says '*saindhavam ānaya*' (bring salt/horse), depending on the context, the word '*saindhava*' may refer to salt in the context of food, or it may refer to horses from Sindh province in a military context.

When words with *vṛtti* (a kind of multi-world expression), such as compound nouns, are present in a sentence, not only *ākāṅkṣā, yogyatā*, *sannidhi*, and *tātparya* contribute to interpretation, but additional conditions also become necessary to establish *anvaya* (semantic connection), thereby enabling verbal cognition (*śābdabodha*) among those compound expressions. There are many discussions in the Indian grammatical tradition (IGT)[3] about the problems and solutions to such situations. Let us now briefly see what *vṛtti* is and how many types of *vṛttis* there are.

1.5. Introduction to vṛttis

In language use, speakers tend to convey a wide range of meanings as economically as possible. To achieve such economy of expression, languages employ systematic grammatical mechanisms, such as suffixation and compounding, through which additional or complex meanings can be encoded within a single linguistic form. This method is called '*vṛtti*' in IGT. Patañjali has defined *vṛtti* in the following manner:

"*parārthābhidhānaṃ vṛttiḥ*" [Patañjali, Kaiyaṭa and Bhaṭṭa 2018, *328*], MBh. I. 364.6–365.14 ad A. 2.1.1[4]

"The expression of the meaning of one word through another is called *vṛtti*". The translation of *śabda* as "word" adopted here is supported by Sastri's discussion of the passage [Sastri 2015, V, *193*].

In this passage, Kaiyaṭa writes, "*parasya śabdasya yo'rthas tasyābhidhānaṁ śabdāntareṇa yatra sā vṛttir-ity-artha* ". That means *vṛtti* is where there is a designation of the meaning of one speech form by another speech form. The term *vṛtti* is a technical term. It denotes a specific formation of sentence structure or word structure. If we want to convey a meaning with a minimum number of words, typically two or more, then *vṛtti* plays a significant role in such contexts. In other words, in a *vṛtti*, semantically compatible multiple words transform into a single word.

For example, let us look at (Ex. 3):

[3] **Indian Grammatical Tradition (IGT)** refers to the Indian classical intellectual tradition concerned with grammatical principles established within disciplines such as **Vyākaraṇa, Nyāya, Mīmāṃsā**, and related śāstric systems.

[4] Mahābhāṣya references follow the Kielhorn-Abhyankar edition and are cited as Volume.Page.Line(s).

(Ex. 3) ***Rāja-puruṣaḥ***

Samāsa-vṛtti	**Vigraha-vākya** (Phrase without *vṛtti*)
Skt: *rāja-puruṣaḥ* **Gls:** royal man *nom. sg* **Eng:** A royal man	**Skt:** *rājñaḥ puruṣaḥ* **Gls:** king *gen. sg*. man *nom. sg* **Eng:** A man of the king

In (3), *rāja-puruṣaḥ* (royal man) is a compound derived from the underlying expression *rājñaḥ puruṣaḥ* (the man of the king). This formation exemplifies a *ṣaṣṭhī-tatpuruṣa samāsa*, a type of compound in the Sanskrit grammatical tradition in which the first member stands in a genitive (*ṣaṣṭhī*) relation to the second, corresponding roughly to an 'X **of** Y' relation in English. In such compounds, the genitive relation is expressed implicitly through compounding rather than overt case marking. That conveys the same meaning as the phrase shown in the *vigraha vākya* column. This is a strategy that makes crucial use of economy in the formation of a sentence. This kind of economy can be seen in all types of *vṛttis*.

Vṛttis are complex word formations. To explain their meaning, grammarians use some phrases called *vigraha vākya*. Varadarājācārya defines *vigraha* as:

"*vṛtty-arthāvabodhakaṃ vākyaṃ vigrahaḥ*" [Varadarājācārya 2004, *172*].

"A phrase that conveys the meaning of a *vṛtti* (such as a compound) is a *vigraha-vākya*".

There are two kinds:

● *Laukika-vigraha-vākya* – a form used in ordinary speech.

● *Alaukika-vigraha-vākya* – a form not used in daily communication but constructed for grammatical analysis.

For example, *rāja-puruṣa* (royal man) is a *samāsa-vṛtti*.

● Its *laukika-vigraha*: *rājñaḥ puruṣaḥ* (man of the king).

● Its *alaukika vigraha*: *rājan+ṅas, puruṣa+su*.

Usually, the *laukika-vigraha* uses the same words that form the *samāsa*. However, sometimes the *vigraha* involves a different word (*asva-pada-vigraha*), producing a *nitya-samāsa* (obligatory compound). In short, a *vṛtti* is a complex word, that is, a single lexical item composed of two word-like elements, whereas its *vigraha-vākya* is the corresponding expanded, non-compounded syntactic expression (e.g., *rājñaḥ puruṣaḥ*).

The five types of *vṛttis* are:

♦ **Kṛt-vṛtti** – *Kṛt-vṛtti* refers to primary (deverbal) derivation formed by the addition of a *kṛt* suffix.

• e.g., *kumbha-kāraḥ* (potter) from *kumbhaṃ karoti* (makes pots) by adding suffix *-aṆ* to √*kṛ* with *kumbha* as object. Here, only *-kāraḥ* is a *kṛt* (deverbal) derivative, while the full form *kumbha-kāraḥ* is an *upapada-samāsa*, formed with *kumbha* as an object-related preverbial element.

♦ **Taddhita-vṛtti** – secondary derivation by adding a *taddhita* suffix to a noun.

• e.g., *dāśarathiḥ* (descendant of Daśaratha) from *daśarathasya apatyam* by adding suffix *-iñ*.

♦ **Samâsa-vṛtti** – compounding two or more words into one for concision.

• e.g., *rāja-puruṣa* from *rājan* + *puruṣa* using *ṣaṣṭhī-tatpuruṣa* compound.

♦ **Ekaœeṣa-vṛtti** – ellipsis of repeated words with identical form.

• e.g., *Rāmaḥ, Rāmaḥ, Rāmaḥ* → *Rāmāḥ.*

• Special rules also allow the ellipsis of different words, e.g., *mātā ca pitā ca* → *pitarau*.

♦ **Sanâdyantadhâtu-vṛtti** – formation of an augmented root by adding certain affixes (*san*, etc.) to a root.

• e.g., *didṛkṣati* (desires to see) from *draṣṭum icchati* by adding suffix *-san* to √*dṛś*.

Modern scholarship has also examined Sanskrit compounds from broader linguistic and theoretical perspectives. George Cardona has emphasised that Pāṇinian compound formation cannot be understood merely as a formal morphological operation, but must be analysed in relation to semantic dependency, grammatical architecture, and usage-based interpretation within the Sanskrit grammatical tradition [Cardona 1997]. Similarly, studies in Indo-European and Sanskrit linguistics by Hans Henrich Hock and others have explored broader questions concerning compositionality, lexicalisation, and the interaction between syntax and morphology in complex linguistic structures [Hock 1991]. Building upon such discussions, Pablo Molina-Muñoz specifically investigates Sanskrit compounds in relation to the architecture of grammar and argues that Sanskrit compounding reflects a complex interaction between morphology, syntax, and semantic structure rather

than a purely mechanical derivational process [Molina-Muñoz 2013]. These studies collectively demonstrate the continuing relevance of Sanskrit grammatical theory to broader linguistic discussions of semantic compositionality, as well as to semantic interpretation and computational representation within the framework of Universal Semantic Representation (USR).

2. Methodology

This study adopts a qualitative, corpus-informed approach to analyse compound structures within the Universal Semantic Representation (USR) framework from a Pāṇinian grammatical perspective. The methodology integrates traditional linguistic theory with computational representation in order to examine the structure and interpretation of compounds, particularly nested vṛttis.

2.1. Data Source

The primary data for this study are drawn from Hindi geography textbooks published by the National Institute of Open Schooling (NIOS). A total of 1,504 sentences were selected from Chapters 1–4, with additional illustrative examples drawn from later chapters where necessary. These texts were chosen due to their high density of compound constructions, especially in technical and descriptive contexts.

2.2. Analytical Framework

The analysis is conducted using the Universal Semantic Representation (USR) framework, which encodes meaning across multiple layers, including conceptual (ConL), dependency (DepR), and constructional (Cons) representations. Each selected sentence is manually analysed to identify semantic units and their relations, with particular attention to compound structures.

2.3. Analytical Procedure

The analysis proceeds through the following steps:

- Identification of compound structures within each sentence,
- Determination of the underlying *vigraha vākya*,
- Application of Pâṇinian constraints to evaluate compound formation,
- Representation of the structure within the Universal Semantic Representation (USR) framework,
- Examination of nested and recursive compounding patterns, and
- Generation of natural language output (NLG) from the USR representation in order to validate the semantic adequacy and interpretability of the analysed structures.

3. Theoretical Discussion

Sāmarthya is necessary for the aforementioned *vṛttis,* as said Pāṇini on the Pā.sū.2.1.1, "*samarthaḥ pada-vidhiḥ*" [Miśra 2015]. We will explain the concept of *sāmarthya* further in Sections 3.3.1 and 3.3.2. Along with *sāmarthya*, certain rules must be followed in special circumstances. What are these rules? Let's clarify now.

3.1. Non-Formation of Vṛtti under External Modifier Expectancy

The formation of *vṛttis* presupposes semantic compatibility (*sāmarthya*) among the elements involved, a requirement articulated by Pāṇini in the rule "*samarthaḥ padavidhiḥ*" [Miśra 2015]. Nevertheless, in later grammatical interpretations, compound formation may still be restricted in certain contexts despite semantic compatibility. Therefore, it is essential to take these prohibitive conditions into account when analysing compounds. Patañjali discusses an additional rule, or condition, that must also be considered in the formation of a *vṛtti.* Let's see:

Rule 1: "*saviśeṣaṇānāṃ vṛttirna vṛttasya vā viśeṣaṇa-yogo na*" [Patañjali, Kaiyaṭa & Bhaṭṭa, No. 2018, *320*].

"*Padas*, that is, inflected words, defined by Pāṇini as *sup-tiṅ-antaṃ padam* [Miśra 2015] (forms ending in nominal or verbal inflection) that exhibit expectancy for modifiers or qualifiers, do not by themselves form a *vṛtti.* Having an expectancy for modifiers/qualifiers does not form a *vṛtti*; also, a component of a *vṛtti* does not relate to other modifiers/qualifiers outside the *vṛtti*".

What is the importance of this principle/rule? Why is it necessary to consider them in *vṛtti* formation? Let's delve into these topics for further discussion on this fundamental principle about a *vṛtti.*

This principle consists of two statements, considered rule 1.a and 1.b.

Rule 1.a: *Saviśeṣaṇānāṃ vṛttir na*

Padas (words) having an expectancy for modifiers/qualifiers do not form a *vṛtti.*

Rule 1.b: *Vṛttasya vā viśeṣaṇa-yogo na*

A component of a *vṛtti* does not relate to other modifiers/qualifiers outside the *vṛtti.*

Rule 1.a states that words with an expectancy for qualifier/s (specific attribute/s) other than those involved in the construction of the *vṛtti* cannot form a *vṛtti*. For instance, look at (Ex. 4).

(Ex. 4) Sanskrit: *ṛddhasya rājñaḥ puruṣaḥ*
Gloss: rich *gen. sg.* king *gen. sg* man *nom.sg*
English: A man of a rich king.

Similar to (Ex. 3), from (Ex. 4), one might wish to derive a *ṣaṣṭhī-tatpuruṣa-samāsa*, that is, a determinative compound in which the non-head constituent bears an implicit genitive relation; '*rāja-puruṣa*' using the words *rājan* (king) and *puruṣa* (man). However, since the word *ṛddha* (rich) functions as a modifier and thus has an expectancy for a head noun, Rule 1.a excludes the formation of a *vṛtti* in this case, rendering *rājan* ineligible for compounding. Consequently, *ṛddhasya rājñaḥ puruṣaḥ* is interpreted as an ordinary syntactic expression rather than as a compound. If 'the rich king's man' meaning is to be conveyed by *samāsa*, then first *ṛddha* and *rājan* need to be combined using a *karmadhāraya-samāsa* and formed as a *ṛddha-rājan*. After that, the word *ṛddha-rājan* should be compounded with *puruṣa* using a *ṣaṣṭhī-tatpuruṣa-samāsa*, that is, a determinative compound in which the non-head constituent bears an implicit genitive relation as a final word '*ṛddha-rāja-puruṣaḥ*' to convey the meaning 'the rich king's man'.

Rule 1.b states that a component of a derived compound may also not have an expectancy for a modifier outside the compound, as shown in (Ex. 5).

(Ex. 5) Sanskrit: **ṛddhasya rāja-puruṣaḥ*
Gloss: rich *gen. sg* royal man *nom. sg*
English: *The rich's royal man.

In (Ex. 5), when the words *rājan* and *puruṣa* are combined by *ṣaṣṭhī-tatpuruṣa-samāsa*, that is, a determinative compound in which the non-head constituent bears an implicit genitive relation, the result comes as a *rāja-puruṣa* (royal man). After the compounding in *rāja-puruṣa* (royal man), if one wishes to add a *viśeṣaṇa* (modifier) to *rājan* (king), restriction of Rule 1.b will be applied. Once this compound word '*rāja-puruṣa*' is formed, the external modifier element, such as *ṛddha* (rich), cannot be accepted within the compound. When a samāsa (compound) is formed, the resulting expression becomes a distinct lexical unit with a specific meaning based on *ekārthībhāva-sāmarthya*, that is, the capacity of the compound to convey a single, integrated meaning (see Section 3 below). If one wishes to connect

some modifier with the compound word, that will connect with the *viśeṣya* (substantive) in the entire distinct meaning of the compound, not with the *viśeṣaṇa* (modifier). This is what Rule 1.b *vṛttasya ca viśeṣaṇa-yogo na* signifies. Therefore, due to Rule 1.b, the modifier *ṛddha* (rich) cannot be correlated with *rājan* (king) within the compound *rājapuruṣa* (royal man).

This Rule 1 is the foundational principle based on the grammatical structures *vṛttis* and *śābda-bodha* (verbal cognition). It emphasises that relational integrity within the *vṛttis* is preserved, and external modifiers cannot modify internal components of the *vṛttis.*

3.2. Exceptions to Rule 1

Rule 1 is broadly applicable, but it has exceptions under specific conditions. In particular examples, external modifiers are mandatory due to the *ākāṅkṣā* of a component of the *vṛttis*. When one element of a *vṛtti* has an expectancy for an external modifier, it cannot convey its meaning properly without the external modifier. In such cases, a *vṛtti can* be formed despite the expectancy for modification outside the components of the *vṛtti.* After creating the *vṛtti*, the external modifier will also be connected with the element of the *vṛtti*.

Relational Words (Sambandhiśabda):

A primary exception arises when a component of a *vṛtti* is a *sambandhi-śabda*, a word inherently relational. For instance, let's see the compound *vṛtti* in (Ex. 6):

(Ex. 6) Sanskrit: *devadattasya guroḥ putraḥ*
Gloss: Devadatta *gen. sg.* guru *gen. sg.* son *nom. sg.*
English: The son of the guru of Devadatta.

In (Ex. 6), by a *ṣaṣṭhī-tatpuruṣa samāsa* between the words' *guru* and *putra*, the compound word '*guru-putra*' is to be formed. But, since the expectancy of '*guru*' is not only with '*putra*' but also with Devadatta*,* a word outside of *vṛtti,* how can the *samāsa-vṛtti* be formed as '*guru-putraḥ*'? Due to Rule 1, this *samāsa* '*guruputraḥ*' cannot be formed.

However, where there is a *niyatā ākāṅkṣā* (permanent expectancy) of the word, like *guru*, the component of *vṛtti*, Rule 1 does not apply as a solution suggested by the traditional scholars as: "*sasaṃbandhika-padārthasya-ekadeśatve'pi bhavaty eva viśeṣaṇānvayaḥ*" (Even when a relationally connected element is a component of a compound, it can nevertheless enter into a relation with an additional modifier)

[Dīkṣita et al. 2022, 1, *466*]. Bhartṛhari has also said the same thing in this context in the following manner:

Rule 1.e:[5] "*sambandhi-śabdaḥ sāpekṣo nityaṃ sarvaḥ samasyate*

vākyavat sā vyapekṣāsya vṛttāv api na hīyate" [Iyer 1974, *148*].

"A *sāpekṣa sambandhi-śabda,* i. e., a word that denotes a relationship, always forms a *vṛtti* despite its meaning-interdependence for a modifier outside the *vṛtti*".

When we form certain compounds, the secondary word may act as a *sambandhi-śabda* (a word denoting a relationship). In such cases, the secondary word can depend on another word outside the expectant compound to communicate its complete meaning. According to Rule 1, a compound cannot be formed in such cases. But, Rule 1.e is an exception to Rule 1. It supports compound formation despite having an expectancy for an external modifier attached to such a secondary word. For instance, as shown in (Ex. 7), the compound '*guru-putraḥ*' must be formed from *guru* and *putra*. G*uru* is the secondary word, and *putra* is the primary word here. The concept of a *guru* is inherently relational because a *guru* exists only in relation to a *śiṣya* (disciple). Someone is called a *guru* when they have a *śiṣya* (disciple). Without a *śiṣya* (disciple), no one is considered a guru. The same condition also applies to the *śiṣya*. Both *śiṣya* and *guru* have *ākāṅkṣā* (expectancy) of each other. Thus, the meaning of the *guru* depends on the idea of a disciple. So, a connection between the *guru* and the *śiṣya* is preserved even when the *guru* is included in a compound, like *guru-putraḥ*.

Since the word *guru* denotes a relational entity, it is a *sākāṅkṣita* word. Since it is *sākāṅkṣita*, Rule 1 cannot be applied here. So, it can take an external modifier from outside the compound. Eventually, *ṣaṣṭhī-tatpuruṣa-samāsa*, where the non-head element bears the implicit genitive case, will be considered between *guru* and *putra*. As a result, '*Devadattasya* ***guru-putraḥ***' will be valid and useful. Let's see in (Ex. 7).

(Ex. 7) Sanskrit: *devadattasya guru-putraḥ*
Gloss: Devadatta *gen. sg* guru-son *nom. sg*
English: The son of Devadatta's guru.

[5] Here, 'e' denotes the exception. So, Rule 1.e means Exception to Rule 1.

Furthermore, Rule 1 does not apply in cases where a *nitya-samāsa* (obligatory compound) is present. For example, let's see in (Ex. 8).

(Ex. 8) Sanskrit: *grāmasya prativṛkṣam*
Gloss: village *gen. sg* each-tree *acc. sg*
English: At every tree in the village.

Here, in (Ex. 8), the form *prativṛkṣam* is derived through an *avyayībhāva* compound formed from *prati* and *vṛkṣa*, conveying the sense of *vīpsā* (pervasion). In such *avyayībhāva* compounds, *vṛkṣa* functions as the *upasarjana* (non-head), being governed by the indeclinable head *prati*. Consequently, when a compound-external modifier such as *grāmasya* is construed as modifying *vṛkṣa*, the resulting construction constitutes a violation of Rule 1, since the modification targets a dependent internal constituent rather than the head of the compound. However, in the context of *nitya-samāsa*, Rule 1 cannot be applied. In (Ex. 8), *prativṛkṣam* is a *nitya-samāsa*, as discussed in Section 1.5. Notice that the *laukika-vigraha* is '*vṛkṣaṁ vṛkṣam*' (tree after tree), which does not have the word '*prati*', whereas the compound has the word '*prati*'. Thus, due to the inapplicability of Rule 1 in *nitya-samāsa*, the *avyayībhāva-samāsa* can be formed using the words *vṛkṣa* and *prati.* Now, let's further explain the causal principle behind Rule 1.b, described in Section 2.

3.3. Optionality of Rule 1

Rule 1 had to be accepted in the context of *vṛtti*. But, if we recognise the *ekārthībhāva-sāmarthya* (described shortly in this Section) for *vṛtti*, then Rule 1 becomes unnecessary. In that case, it is crucial to understand the principle of *ekārthībhāva-sāmarthya.* What role does it play in *vṛttis?* Let us see.

3.3.1. Ekārthībhāva-sāmarthya (Single Integrated Meaning)

In Pāṇinian grammar, '*sāmarthya*' is a technical term that is classified into two categories:

1. *Ekârthîbhâva-sāmarthya*
2. *Vyapekṣābhāva-sāmarthya*

Let us first explain *ekārthībhāva-sāmarthya*. The notion of *ekārthībhāva-sāmarthya* is discussed already in Patañjali's *Mahābhāṣya* [Patañjali, Kaiyaṭa, Bhaṭṭa 2018], where semantic unity is treated as the basis of compound formation. During the explanation of the rule p.2.1.1 *samarthaḥ padavidhiḥ* [Miśra 2015], Vāsudeva Dīkṣita defines *ekārthībhāva-sāmarthya* in the following manner:

"*prakriyādaśāyāṃ pṛthagarthavatvena gṛhītānāṃ padānāṃ samudāyaśaktyāviśiṣṭaikārthapratipādakatvam ekārthībhāva-sāmarthyam*" [Dīkṣita et al. 2022, *467*].

"The *sāmarthya* (capability) of shifting from multiple meanings into a single meaning is called *ekārthībhāva-sāmarthya*".

Ekārthībhāva-sāmarthya arises in formations where padas enter into a close morphosyntactic relation (*pada-kārya*), as in the case of *vṛttis*. For example, in (Ex. 3), when *rājan* and *puruṣa* combine to form the compound *rāja-puruṣaḥ*, the two stems retain their respective meanings king and man, just as they do in the corresponding syntactic construction *rājñaḥ puruṣaḥ* (a man of the king). The crucial difference between the compound and the noun phrase does not lie in a contrast between multiple versus single meanings, but rather in the grammatical status of the resulting form. In compounding, the two stems are unified into a single lexical item, with the case endings of the constituent padas being zero-replaced, yielding a word-level unit. This capacity of morphosyntactic unification into a single lexeme is what is meant by *ekārthībhāva-sāmarthya*. The word *rāja-puruṣa*, arising from this capacity, conveys a specific meaning: the royal man, due to the *ekārthībhāva-sāmarthya*.

3.3.2. Vyapekṣābhāva-sāmarthya (Meaning-Interdependence)

The mutual connection of words in a sentence based on their expectancy, compatibility, and proximity is called *vyapekṣābhāva-sāmarthya. Vāsudeva Dīkṣita* defined the *vyapekṣābhāva-sāmarthya* in the following manner:

"*svārthaparyavasāyināṃ padānām ākāṅkṣādivaśāt yaḥ paraspararaṃ sambandhaḥ sā vyapekṣā*" [Dīkṣita et al. 2022, *467*].

"*Vyapekṣābhāva-sāmarthya* refers to the capability of *padas* to have a mutual relationship and relevance in sentences due to their *ākāṅkṣā, yogyatā*, and *sannidhi*".

For example, in (Ex. 1), the words *Rāma, gṛhaṃ,* and *gacchati* are connected with each other according to their *ākāṅkṣā* (expectancy), *yogyatā* (Meaning Congruity), and *sannidhi* (Contiguity). This connection is known as a *vyapekṣābhāva-sāmarthya,* which conveys the whole meaning of the sentence.

The distinction between *vyapekṣābhāva sāmarthya* and *ekārthībhāva sāmarthya* can already be traced to Patañjali's discussion of *samarthaḥ padavidhiḥ* in the *Mahābhāṣya* [Patañjali, Kaiyaṭa, Bhaṭṭa 2018]. There, Patañjali observes that if *sāmarthya* is interpreted as

ekārthībhāva, the rule successfully accounts for *samāsa*, whereas *vibhakti-vidhāna* and *parāṅgavadbhāva* remain unexplained. Conversely, if *sāmarthya* is interpreted as *vyapekṣā*, it accounts for *vibhakti-vidhāna* and *parāṅgavadbhāva*, but not for *samāsa*[6]. *This discussion establishes a distinction between the semantic unity characteristic of compounds and the mutual expectancy that characterises syntactically related words. Building on this foundation, later Pāṇinian grammarians associated vyapekṣābhāva sāmarthya* primarily with sentences and *ekārthībhāva sāmarthya* with *vṛtti* [Bhaṭṭojidīkṣita et al. 2022, *467*]. According to the theory of *ekārthībhāva sāmarthya*, the constituents of a compound function as a single word, and their meaning is expressed in an integrated manner rather than as the sum of independently functioning *padas*. From this standpoint, the compound is analysed in terms of *jahatsvārthā-vṛtti*: a type of *vṛtti* in which the constituents relinquish their independent denotative meanings (*svārtha*) and yield a single integrated compound meaning. This position is upheld by grammarians such as Bhaṭṭojī Dīkṣita, Nageśa Bhaṭṭā and others [Ashtadhyayi 2026].

However, some grammarians, notably Nāgeśa Bhaṭṭa, as stated, "*jahatsvārthā tu tatraiva yatra rūḍher virodhinī*" (*jahatsvārthā-vṛtti* is to be assumed only where accepting the constituent meanings leads to a contradiction with the conventional sense), do not generally accept *jahatsvārthā-vṛtti* in compounds and similar constructions[7]

[6] "*tatraikārthībhāve sāmarthye'dhikāre ca sati samāsa ekaḥ saṃgṛhīto bhavati| vibhaktividhānaṃ parāṅgavadbhāvaścāsaṃgṛhītaḥ || vyapekṣāyāṃ punaḥ sārmathye'dhikāre ca sati vibhaktividhānaṃ parāṅgavadbhāvaśca saṃgṛhīto bhavati| samāsastveko'saṃgṛhītaḥ||*". See [Patañjali, Kaiyaṭa, Bhaṭṭa 2018] for more information.

[7] "*nanu viśiṣṭa-śakta-svīkāre paṅkaja-padāt avayavārtha-pratītir mā bhūt samudāya-śaktyaiva kamala-padavat puṣpa-viśeṣa-pratyayaḥ syād iti cen na,* ***jahatsvārthā tu tatraiva yatra rūḍhir virodhinī*** *iti abhiyuktokteḥ avayavārtha-saṃvalita-samudāyārthe padme śakti-svīkārāt*". That means:

Objection: If, in accordance with *ekārthībhāva-sāmarthya* in compounds, a *viśiṣṭa-śakti* is accepted, then in the word *paṅkaja* (that which is born in mud), there should be no cognition of the constituent meanings (*paṅka* 'mud' and *ja* 'born'). Instead, only the conventional meaning 'lotus' should be apprehended through that specific power alone.

Reply: This is not the case. *Jahatsvārthā-vṛtti* is accepted only where the etymological meaning (*yaugikārtha*) conflicts with the conventional meaning

[Ashtadhyayi 2026]. They argue instead that *jahatsvārthā-vṛtti* is admitted only in cases where the retention of constituent meanings would conflict with an established or conventional meaning (*rūḍhi*). Otherwise, the default assumption is *ajahatsvārthā-vṛtti:* a type of compounding in which each constituent retains its independent denotative meaning (*svārtha*) and together they produce the meaning of the compound. When *ekārthībhāva-sāmarthya* is adopted in the sense of *jahatsvārthā-vṛtti*, the compound is treated as a single semantic unit (without being subsumed into the whole); since its constituents no longer operate independently, an external modifier cannot be construed as modifying any internal component. According to those who accept *ajahatsvārthā*, the meaning of the constituents in a compound is not relinquished; however, since the compound meaning is specific as *ekārthībhāva*, the meanings of the constituents become partial aspects (*padārthaikadeśa*) of the compound meaning. Because they are only partial aspects, in accordance with the principle "*padārthaḥ padārthena anvēti, na tu padārtha-ekadeśena*", the head meaning (*viśeṣya*)[8] denoted by one word is connected with the head meaning (*viśeṣya*) denoted by another word, not with only a part of that meaning, the possibility of an external modifier attaching to the compound is ruled out.

Since *ekārthībhāva-sāmarthya* is accepted in *vṛtti*, in (Ex. 5), the word *rāja-puruṣa* signifies the complete meaning 'a royal man' (the man related to the king). According to the theory of *ekārthībhāva sāmarthya*, the compound *rājapuruṣa* expresses a single integrated meaning. From the standpoint of *ajahatsvārthā-vṛtti*, the constituent meanings of *rājan* (king) and *puruṣa* (man) are not completely abandoned; rather, they are subsumed into the unified meaning of the compound and are not apprehended as independently functioning meanings. In contrast, according to the theory of *jahatsvārthā-vṛtti*, the constituents relinquish their individual meanings and only the compound meaning is conveyed. So, the modifier '*ṛddhasya*' (of the rich) cannot relate to the *rājan.* So, in this scenario, Rule 1 becomes unnecessary to deny the connection of the modifier (outside of *vṛtti*) with

(*rūḍhyartha*). Where there is no such conflict, *ajahatsvārthā-vṛtti* is the default. Accordingly, in words like *paṅkaja*, since the etymological meaning does not oppose the conventional sense 'lotus', the constituent meanings are also apprehended.

[8] Here, *padārtha* refers to the *viśeṣya* (head) of the meaning.

vṛtti. With this perspective, Kauṇḍa Bhaṭṭa mentioned this idea in his text *Vaiyākaraṇa-bhūṣaṇa-sāraḥ* in the following manner:

"*samāse khalu bhinnaiva śaktiḥ paṅkaja-śabdavat*
bahūnāṃ vṛtti-dharmāṇāṃ vacanair eva sādhane
syān mahad gauravaṃ tasmād ekārthībhāva āśritaḥ" [Pañcholi 2011, *277*].

"Just like the word *paṅkaja* (lotus), in the compound*, ekārthībhāva-sāmarthya* should be considered. Instead of making various rules and theories, such as Rule 1, etc."

This analysis motivates the postulation of a specialised semantic potency (*viśiṣṭa-śakti*) operative in compounds. For example, the etymological derivation of the word *paṅkaja* is *paṅke jātam* (that which is born in mud). Within the Pāṇinian grammatical framework, *paṅkaja* can be analysed as an *upapada-samāsa*, where *paṅka* functions as the *upapada* and '*ja*' is a *kṛt*-derivative formed from the root √*jan*. However, in actual linguistic usage, the word *paṅkaja* does not denote all entities that are born in mud; rather, it conventionally denotes only *padma* (lotus). In order to exclude other potential interpretations, such as objects merely produced in mud, it becomes necessary to assume that *paṅkaja* operates with a specific *ekārthībhāva*-based semantic force, by virtue of which the general meaning of 'being born in mud' (*paṅka-jani-kartṛtva*) is relinquished, and only the determinate sense 'lotus' (*padmatva-viśiṣṭaḥ padmaḥ*) is conveyed. Thus, in general, in compounds (*samāsa*), *ekārthībhāva-sāmarthya* is assumed to account for the establishment of a specific meaning, as in *paṅkaja*. From the perspective of modern linguistics, *paṅkaja* may be regarded as a lexicalised or non-compositional compound, since the meaning of the whole compound 'lotus' cannot be straightforwardly derived from the meanings of its constituents ('mud' and 'born'). In this respect, the traditional notion of *jahatsvārthā-vṛtti* provides an account of a phenomenon that modern linguistic scholarship describes as lexicalisation and non-compositionality [Lowe 2015]. For further discussion, see [Pañcholi 2011, *277–278*].

4. Analysis of Compound Structure in Universal Semantic Representation (USR)

Pāṇinian grammar, though originally written for Sanskrit, discusses language-independent principles [Kulkarni 2007]. These principles discuss semantics, *kāraka* and *kāraketara* relations (dependency),

compound formations, etc. This paper primarily focuses on the principles of compound formation in Pāṇinian grammar and how they can help capture the semantics of compounds in languages other than Sanskrit. ***Aikapadya*** (unification into a single lexical unit), ***aikasvarya*** (assignment of a single accent), and related morphosyntactic properties of compounds are distinctive features of Sanskrit. In contrast, many other Indian and non-Indian languages do not necessarily represent compound expressions as a single orthographic word. Although such expressions may constitute a single lexical unit, their components may remain graphically separated according to the orthographic conventions of the language. Sanskrit, by contrast, typically represents compounds as a single continuous orthographic form. A language may employ multiple orthographic conventions in representing the elements of a compound. For example, in Hindi, compound words may be written in three different ways: with a hyphen, as in ***bhū-bhāga*** (land area), as a single word, like ***bhūtala*** (surface), or as separate words, like ***rāhata sāmagrī*** (relief material). Similarly, English also follows these writing styles for writing compound words. For example, as a single word like **milkman** (a person who delivers milk), hyphenated as in **milk-fever** (a disease caused by a lack of the calcium contained in milk), and with a space as in **milk bottle** (a bottle used to contain milk), etc. [Sukhada 2017, *30*]. Due to this, when compounds involve more than two words or are nested, identifying compounds and the connection between their components (i.e., which component is connected to which) becomes more challenging for machines. However, despite this non-uniformity in writing style, Rule 1 can help achieve a uniform understanding of the semantics of compounds by providing specific criteria for the connection of compound components and their formation, as explained in Section 2. Therefore, we decided to apply this theory to Universal Semantic Representation (USR) [Garg et al. 2023, *14–22*] to provide a uniform representation of compound semantics.

4.1. Universal Semantic Representation

The Universal Semantic Representation (USR) is a framework for meaning representation inspired by the Indian Grammatical Tradition (IGT). USR aims to capture the speaker's or author's intention. In communication (whether oral or written), one uses various linguistic

expressions in sentences to convey their thoughts to the listener or reader. Such information is represented within the USR at three levels. The USR framework represents this information in a computationally tractable format, enabling formal semantic analysis and supporting applications in natural language processing. Each level contains one or more layers, also called rows, to capture different kinds of linguistic and semantic information. These three levels are as follows:[9]

1. **Lexico-conceptual Level:** This level captures conceptual information typically expressed through atomic words, multiword expressions, or derived words. Currently, it contains information organised under the following labels within the USR [Garg et al. 2023, *14*].

a. **ConL** or **Con(cept) L(abel):** The concepts (semantic constructs), like entities, events, and their modifiers, are represented under the concept label heading. The concept labels also include complex entities such as compounds, temporal expressions consisting of multiple subphases, coordinating conjunctions, and disjunctions. The concepts specific to these entities are represented within brackets [].

b. **Index:** Each concept is given an index number to uniquely identify each concept label to help mark the head-dependency, coreferencing, and compositionality among members of concepts, represented under the Syntactico-Semantic Label.

c. **SemCat** or **Sem(antic) Cat(egory):** In this layer, the semantic categories of the concepts, such as a person, organisation, place, named entities (NE), time, number, animacy, etc., are specified.

d. **MoSem** or **Mo(rpho) Sem(antic Information):** This layer captures the morphological information of the concepts, such as degree of comparison, causation, etc., typically expressed through derivational morphology applied to the root words.

2. **Syntactico-Semantic or Propositional Level:** At this level, relational information, commonly conveyed through *kāraka-vibhakti*, non-*kāraka-vibhakti*, compounding, conjunction, etc. [Garg et al. 2023, *14–15*] is captured. It includes the following layers:

a. **DepR** or **Dep(endency) R(elations):** This label defines the relationship between the head and its dependent/s by specifying the

[9] In the future, USR will represent the scope-level information as well. Research on the scope is in progress at this level.

index of the head. Here, head-dependent relations in terms of **kāraka** and **non-kāraka** relations are represented.

b. **Cons** or **Cons(truction):** USR defines Constructions as semantic frames expressed through linguistic units larger than a word but smaller than a sentence, such as compounds, temporal expressions, conjunctions, etc. This type of semantic information is captured under the Construction layer.

3. **Discourse Level:** This level captures inter-sentential and intra-sentential information such as cohesion and coherence, and the speaker's perspective as emphasising, distinguishing, etc. [Garg et al. 2023, *14–15*]. It includes the following layers:

a. **DiscE** or **Dis(course) E(lements):** This layer captures the discourse-level connections that ensure cohesion and coherence in communication. It focuses on discourse connective relations and pronominal coreference across or within sentences.

b. **SpekV** or **Speak(er's) V(iew):** The speaker's view layer represents the speaker's perspective as emphasis, respect, etc., conveyed in the discourse.

Now, let us see an example of Universal Semantic Representation (USR) through a sentence (Ex. 9). In this paper, Hindi Geography data from NIOS (National Institute of Open Schooling) have been used for compound analysis, mainly from Chapters 1–4. Examples from chapters 10 and 11 have also been provided for illustrative purposes.

(Ex. 9) *bāḍha yā bhūkampa ke bāda sabhī vyaktiyoṃ ko rāhata sāmagrī kī jarūrata hotī hai* [Geography (316) syllabus, *7*].

Hin:*bāḍha yā bhūkampa ke bāda sabhī*

Gls: flood *nom. f. sg.* or *disjunct* earthquake *gen. m. sg.* after *post. loc.* all *quant*

vyaktiyoṃ korāhata sāmagrī kī jarūrata hotī hai.

person *dat. m. pl*. relief material *gen. f. sg.* necessity *nom. f. sg.* be *pres. 3. sg*.

Eng: After a flood or earthquake, all persons require relief materials.

Table 1. Universal Semantic Representation of the (Ex. 9)

<sent_id= Geo_nios_1ch_0100>							
ConL	**Ind**	**SemCat**	**MoSem**	**DepR**	**DiscR**	**SpekV**	**Cons**
bāḍha_1	1	–	–	–	–	–	3:op1
bhūkampa_1	2	–	–	–	–	–	3:op2
[disjunct_1]	3	–	–	4:rkl	–	–	–
bāda_1	4	–	–	11:k7t	–	–	–
saba_1	5	–	–	6:quant	–	hī_1	–
vyakti_2	6	anim	pl	11:k4a	–	–	–
rāhata_2	7	–	–	–	–	–	**9:sādhya**
sāmagrī_2	8	–	–	–	–	–	**9:sādhana**
[4-tat_1]	9	–	–	10:r6	–	–	–
jarurata_1	10	–	–	11:k1	–	–	–
ho_1-tā_hai_1	11	–	–	0:main	–	–	–
</sent_id>							

For the convenience of readers unfamiliar with the USR framework, a glossary of the abbreviations and Sanskrit grammatical terms appearing in the tables is provided in the Appendix (Table 9).

Table 1 presents the Universal Semantic Representation (USR) of the Hindi sentence (Ex. 9), illustrating how meaning is encoded across the lexico-conceptual, syntactico-semantic, and discourse levels. Geo_nios_1ch_0100 is the sentence ID of example (Ex. 9). Such sentence IDs are assigned in the USR to enable inter-sentential reference and cross-referencing across the dataset. The **ConL** column lists the conceptual units invoked in the sentence, including entities (*bāḍha* 'flood', *bhūkampa* 'earthquake', *vyakti* 'people'), modifiers (*sabhī* 'all'), predicates (*jarurata* 'need', *ho* 'be'), and complex constructions such as compounds and discourse operators, which are represented within square brackets (e.g. [4-tat_1], [disjunct_1]). Each concept is assigned a unique **Index (Ind)**, which serves as a reference point for encoding dependency relations and constructional membership within the USR. The **SemCat** column specifies semantic categories where relevant; for instance, *vyakti* is marked as animate and plural. The **MoSem** layer captures morpho-semantic information, though no overt morpho-semantic distinctions are required for most items in this sentence. At the syntactico-semantic level, the **DepR** column encodes *kāraka* and non-*kāraka* relations by indicating the head index

and the type of relation, such as '*sabhī*' functioning as a quantifier (5:quant) of '*vyakti*', or *vyakti* bearing the k4a relation to the event of need. The **Cons** column represents constructional information, including the relational role of the compound *rāhata sāmagrī*, analysed as [4-tat_1] at the concept level. As highlighted in Table 1 [4-tat_1] indicates a *caturthī tatpuruṣa-samāsa*, that is, a determinative compound in which a dative ('for') relation holds between the compound members *rāhata_2* 'relief' and *sāmagrī_2* 'material'. The compound *rāhata-sāmagrī* is semantically equivalent to the paraphrase *rāhatāya sāmagrī* ('material **for** relief'), where the dative relation applies to the *upasarjana* (non-head) *rāhata*, and *sāmagrī* functions as the head. Accordingly, at the Construction level in the USR, *rāhata_2* and *sāmagrī_2* are marked with ***sādhya*** (purpose) and ***sādhana*** (means) relations[10], respectively, reflecting the purposive relation encoded by the compound. The **SpekV** column remains empty as no explicit speaker-oriented perspective is encoded. Taken together, the table demonstrates how USR systematically integrates conceptual content, grammatical relations, compounding, and discourse structure to represent the intended meaning of the sentence.

4.2. Application of Rule 1 in USR for Indian Language

In case a component of the *vṛtti* that acts as a modifier to the *vṛtti* has a modifier outside the *vṛtti*, it cannot form a *vṛtti* due to Rule 1. So, initially, an external modifier and component of the *vṛtti* form a compound. After that, this compound can form another *vṛtti*, creating a nested *vṛtti*. Each *vṛtti* has the potential to contain nested *vṛtti (vṛtti-garbhaka vṛtti)*. Following this idea, Rule 1 will be applied to mark the compound and other vṛttis within the Universal Semantic Representation (USR). There can be multiple layers of *vṛttis*, where a *vṛtti* exists within another *vṛtti*, such as compound, and so on. Each *vṛttis* can be nested. However, in the NIOS data analysed for this paper, only

[10] In the Universal Semantic Representation (USR), relations may be labelled using either English terms (like 'mod', 'head', 'quant') or Sanskrit terms (like '*sādhya*', '*sādhana*', '*viṣaya*', '*viṣayi*', etc.). The choice of terminology does not affect the underlying representation, as USR's primary focus is the semantic representation of meaning rather than the language of the labels. In the USR framework, each relational term is mapped to its equivalent term that conveys the same meaning, ensuring consistent semantic representation across labelling languages.

specific types of nested *vṛttis*, such as compounds, have been identified. There are a few types of nested *vṛtti*, as shown in Table 2 below.

Table 2. Types of nested vṛtti

Samāsa-vṛtti-garbhaka samāsa-vṛtti	*Samāsa-vṛtti-garbhaka anya-vṛtti*
Karmadhāraya-garbhaka tatpuruṣa-samāsa	*Tatpuruṣa-garbhaka taddhita-vṛtti*
Tatpuruṣa-garbhaka tatpuruṣa-samāsa	
Dvandva-garbhaka tatpuruṣa-samāsa	

Now, let us explain these nested *vṛttis* with examples, respectively.

4.2.1. Samāsa-vṛtti-garbhaka samāsa-vṛtti (Nested Compounds)

Samāsa-vṛtti-garbhaka samāsa-vṛtti is a type of compound formation in which an internally formed compound (*samāsa-vṛtti*) is embedded within another compound and functions as a single *samāsta-pada* for further compounding, thereby exhibiting the recursive nature of samāsa formation. Such constructions are not uniform but encompass multiple types, depending on the nature and interaction of the internal and external samāsa, which are discussed in further detail.

- **Karmadhâraya-garbhaka tatpuruṣa-samāsa (Appositional Compound with Determinative Compound)**

A *Karmadhāraya-garbhaka tatpuruṣa-samāsa* denotes a multilayered grammatical process where an initial pair of constituents undergoes a *karmadhāraya* transformation (appositional/descriptive compounding) to form a unified base. This base then functions as a single *samāsta-pada* (compound word) to relate to a third constituent through a *tatpuruṣa* (determinative) process. For example, let us examine (Ex. 10), (Ex. 11), or (Ex. 12).

(Ex. 10) *prārambhika vidvān varṇanātmaka bhūgolavettā the* [Geography (316)].

Hin: *prārambhika vidvān varṇanātmaka bhūgola-vettā the.*

Gls: earlier *adj.* scholar *nom. m. pl.* descriptive adj. geographer *nom. m.pl.* be *past. 3. pl.*

Eng: The earlier scholars were descriptive geographers.

In (Ex. 10), the word *bhūgola-vettā* is formed by a *ṣaṣṭhī-tatpuruṣa* compound, i. e., a determinative compound in which the semantic relation between the members corresponds to a genitive (‘of’) relation in the underlying paraphrase, between *bhūgola* ‘geography’ and *vettā* ‘knower’. The word *varṇanātmaka* (descriptive) is a modifier of *bhūgola* (geography). However, since *bhūgola-vettā* (geographer) is a

compound word, the modifier (outside of the compound) *varṇanātmaka* (descriptive) cannot be connected with *bhūgola* due to Rule 1. Therefore, it is essential to form the first '*varṇanātmaka-bhūgola*' using a *karmadhāraya-samāsa* between the words *varṇanātmaka* and *bhūgola*. After that, '*varṇanātmaka-bhūgola-vettā*' will be formed using *ṣaṣṭhī-tatpuruṣa-samāsa* (a determinative compound where the non-head element bears implicit genitive case) between the words *varṇanātmaka-bhūgola* and *vettā.* This approach provides the correct semantic interpretation. This is illustrated in Figure 1.

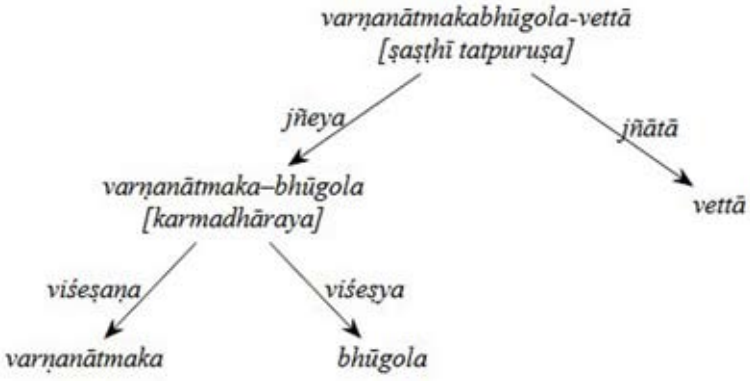

Fig. 1. compound structure of the (Ex. 10)

Now, let us see the USR in Table 3:

Table 3. Universal Semantic Representation of the (Ex. 10)

<sent_id= Geo_nios_1ch_0011>							
ConL	**Index**	**SemCat**	**MoSem**	**DepR**	**DisE**	**SpeakV**	**Cons**
prāraṃbhika_6	1	–	–	2:mod	–	–	–
vidvāna_6	2	anim	pl	8:k1	–	–	–
varṇanātma-ka_1	3	–	–	–	–	–	**5:viśeṣaṇa**
bhūgola_1	4	–	–	–	–	–	**5:viśeṣya**
[karmad-hāraya_1]	5	–	–	–	–	–	**7:jñeya**
vettā_1	6	–	–	–	–	–	**7:jñātā**
[6-tat_1]	7	–	pl	8:k1s	–	–	–
hai_1-past	8	–	–	0:main	–	–	–
</sent_id>							

In Table 3, **[karmadhāraya_1]** denotes a *karmadhāraya-samāsa* between the varṇanātmaka_1 and bhūgola_1. In the construction, **mod** and **head** relations are assigned to varṇanātmaka_1 and bhūgola_1, respectively, with respect to index **5 [karmadhāraya_1].** Similarly, **[6-tat_1]** denotes a *ṣaṣṭhī-tatpuruṣa-samāsa* (a determinative compound where the non-head element bears implicit genitive case) between **[karmadhāraya_1]** (*varṇanātmaka-bhūgola*) and vettā_1. So, in the construction label, ***jñeya*** and ***jñātā*** relations are given to [**karmadhāraya_1]** (*varṇanātmaka-bhūgola*) and *vettā*, respectively.

(Ex. 11) *mānacitroṃ ko banāne ke lie aṃkīya bhaugolika sūcanā tantra naye upakaraṇa haiṃ* [Geography (316)].

Hin: *mānacitroṃ ko banāne ke lie aṃkīya bhaugolika*

Gls: map *acc. m. pl.* make *inf. purpose* digital *adj.* geographical *adj.*

sūcanā tantra naye upakaraṇa haiṃ

information *nom. f. pl.* system *nom. m. pl.* new *adj.* tool *nom. m. pl.* be *pres. 3 pl.*

Eng: Digital geographical information systems are new tools for making maps.

In (Ex. 11), *sūcanā-tantra* ‘information system’ is analysed as a *ṣaṣṭhī-tatpuruṣa* (genitive) compound, formed from *sūcanā* ‘information’ and *tantra* ‘system’. The genitive relation corresponds to an English of-phrase (i.e., ‘system of information’) and reflects the role of the non-head element within the compound, rather than the case of that non-head element. The noun *sūcanā* ‘information’ is further modified by *bhaugolika* ‘geographical’. In this situation, Rule 1 prevents a direct compound formation between *sūcanā* ‘information’ and *tantra* ‘system’. Consequently, the compound *bhaugolika-sūcanā* ‘geographical information’ must first be formed as a *karmadhāraya* compound between *bhaugolika* ‘geographical’ and *sūcanā* ‘information’. In the next step, a *ṣaṣṭhī-tatpuruṣa* (genitive) compound is formed between *bhaugolika-sūcanā* and *tantra* ‘system’, yielding *bhaugolika-sūcanā-tantra* ‘geographical information system’. Semantically, this corresponds to an English paraphrase such as ‘a system for geographical information’. See Figure 2.

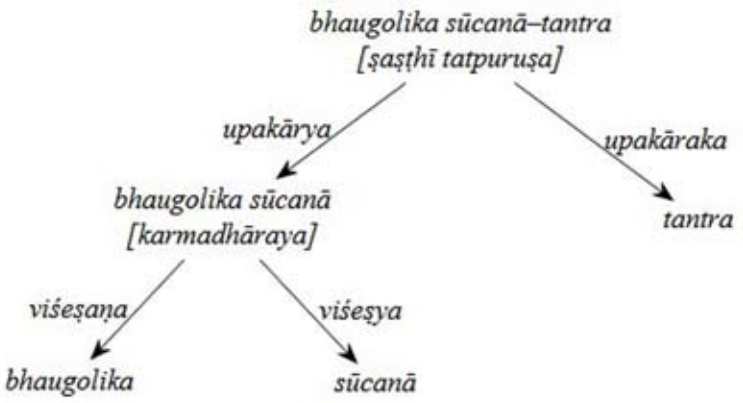

Fig. 2. The compound structure in (Ex. 11)

The relations used in all figures and USR representations are defined and explained in the Appendix. Let's see USR in Table 4:

Table 4. Universal Semantic Representation of the (Ex. 11)

<sent_id= Geo_nios_1ch_0048>							
ConL	**Index**	**SemCat**	**MoSem**	**DepR**	**DisE**	**SpeakV**	**Cons**
mānacitra_1	1	–	pl	2:k2	–	–	–
banā_1	2	–	–	11:rt	–	–	–
aṃkīya_3	3	–	–	7:mod	–	–	–
bhaugolika_1	4	–	–	–	–	–	**6:mod**
sūcanā_1	5	–	–	–	–	–	**6:head**
[karmad-hāraya_1]	6	–	–	–	–	–	**8:upakā-rya**
tantra_1	7	–	–	–	–	–	**8:upakā-raka**
[6-tat_1]	8	–	–	11:k1	–	–	–
nayā_1	9	–	–	10:mod	–	–	–
upakaraṇa_1	10	–	–	11:k1s	–	–	–
hai_1-pres	11	–	–	0:main	–	–	–
</sent_id>							

In Table 4, **[karmadhāraya_1]** denotes a *karmadhāraya-samāsa* between the words *bhaugolika* and *sūcanā*. In the given construction,

the mod and head relations are assigned, respectively, to the bhaugolika and sūcanā for reference under index **6 [karmadhāraya_1]**. Similarly, **[6-tat_1]** indicates *ṣaṣṭhī-tatpuruṣa-samāsa* (a determinative compound where the non-head element bears implicit genitive case) between **[karmadhāraya_1]** (*bhaugolika sūcanā*) and *tantra*. So, ***upakārya*** and ***upakāraka*** relations are given in the construction to [karmadhāraya_1] (i. e., *bhaugolika sūcanā*) and *tantra*, respectively. Because the system supports, processes, and manages the information. To know the full form or definition of these relational tags, see the appendix (table 9) at the end.

- **tatpuruṣa-garbhaka tatpuruṣa-samāsa (Determinative with Determinative Compound)**

A *Tatpuruṣa-garbhaka tatpuruṣa-samāsa* denotes a recursive determinative compound where the internal structure (the *garbha*) is itself a *tatpuruṣa* compound. In this formation, an initial pair of constituents first undergoes a *tatpuruṣa* transformation to form a unified base, which then functions as a single *samāsta-pada* to combine with a third constituent in a secondary *tatpuruṣa* process. For instance, let us see (Ex. 12):

(Ex. 12) ***bhū ākṛti vijñāna****, jalavāyu vijñāna, samudra vijñāna, mṛdā aura jaiva bhūgola* [Geography (316)].

Hin: *bhū ākṛti vijñāna, jalavāyu vijñāna, samudra vijñāna, mṛdā aura jaiva bhūgola*

Gls: earth *gen. f. sg.* shape *gen. f. sg.* science *nom. m. sg*. climate *gen. m. sg.* science *nom. m. sg.* ocean *gen. m. sg*. science *nom. m. sg.* soil *nom. m. sg.* and *conj.* bio *adj.* geography *nom. m. sg.*

Eng: geomorphology, climatology, oceanography, soil and biogeography.

As highlighted in (Ex. 12), *bhū ākṛti vijñāna* (geomorphology) is a compound word. Here, *bhū* (earth) functions as a genitive modifier of the head *ākṛti* (shape), and *bhū ākṛti* (shape of the earth) also functions as a genitive modifier of the head '*vijñāna*' (science). So *bhū ākṛti vijñāna* includes two compounds. Initially, *bhū ākṛti* is formed by the *ṣaṣṭhī-tatpuruṣa-samāsa* of *bhū* and *ākṛti*. After that, using *ṣaṣṭhī-tatpuruṣa-samāsa* again between the words *bhūākṛti* and *vijñāna*, '*bhū ākṛti vijñāna*' will be completely formed.

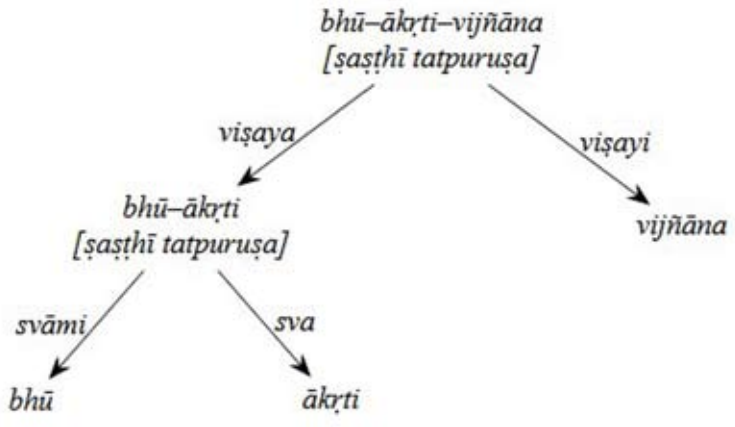

Fig. 3. The compound structure in (Ex. 12)

Now, let us see USR in Table 6:

Table 6. Universal Semantic Representation of the (Ex. 12)

<sent_id= Geo_nios_1ch_0188>							
ConL	**Index**	**SemCat**	**MoSem**	**DepR**	**DisE**	**SpeakV**	**Cons**
bhū_1	1	–	–	–	–	–	**3:sva**
ākṛti_1	2	–	–	–	–	–	**3:svāmi**
[6-tat_1]	3	–	–	–	–	–	**5:viṣaya**
vijñāna_1	4	–	–	–	–	–	**5:viṣayi**
[6-tat_2]	5	–	–	–	–	–	17:op1
jalavāyu_1	6	–	–	–	–	–	8:viṣaya
vijñāna_1	7	–	–	–	–	–	8:viṣayi
[6-tat_3]	8	–	–	–	–	–	17:op2
samudra_1	9	–	–	–	–	–	11:viṣaya
vijñāna_1	10	–	–	–	–	–	11:viṣayi
[6-tat_4]	11	–	–	–	–	–	17:op3
mṛdā_1	12	–	–	–	–	–	14:mod
bhūgola_1	13	–	–	–	–	–	14:head
[6-tat_5]	14	–	–	–	–	–	17:op4
jaiva_1	15	–	–	16:mod	–	–	–

bhūgola_1	16	–	–	–	–	–	17:op5
[conj_1]	17	–	–	8:re	–	–	–
[6-tat_6]	18	–	–	–	–	–	17:op6
</sent_id>							

In Table 6, **[6-tat_1]** denotes a *ṣaṣthī-tatpuruṣa-samāsa* between the bhū_1 and ākṛti_1. **[6-tat_2]** denotes a *ṣaṣthī tatpuruṣa samāsa* between the **[6-tat_1]** *(bhū-ākṛti)* and *vijñāna*. The relationships among the components of the compound are given in the construction.

- **Dvandva-garbhaka tatpuruṣa-samāsa (Copulative Compound with Determinative Compound)**

A *Dvandva-garbhaka tatpuruṣa-samāsa* denotes a multi-layered formation where an initial set of constituents undergoes a *dvandva* transformation (copulative compounding) to create a unified base, which then functions as a single *samāsta-pada* to relate to a subsequent word through a *tatpuruṣa* process. For example, consider the following (Ex. 13).

(Ex. 13) *Anyā mahatvapūrṇa vicārakoṃ ne bhūgola ko mānava-paryāvaraṇa-antarsambandhoṃ ke rūpa meṃ paribhāṣita kiyā hai* [Geography (316)].

Hin: *anya mahatvapūrṇa vicārakoṃ ne bhūgola ko mānava- paryāvaraṇa- antarsambandhoṃ ke rūpa meṃ paribhāṣita kiyā hai.*

Gls: other *mod.* important *adj.* thinker *nom. m. pl.* geography *acc. m. sg.* human *nom. m. sg.* environment *nom. m. sg.* -interrelation *nom. m. pl.* as form *loc. m. sg.* define *perf. 3. sg.*

Eng: Other important thinkers have defined the geography as a form of human-environment interrelation.

In (Ex. 13), the word '*mānava-paryāvaraṇa antarsambandha*' is compounded. There is a *Dvandva-garbhaka tatpuruṣa-samāsa.* Initially, the '*mānava-paryāvaraṇa*' word is formed using *dvandva-samāsa* between the word *mānava* and *paryāvaraṇa*. After that, '*mānava-paryāvaraṇa*' is connected with *antarsambandha* to form '*mānava-paryāvaraṇa antarsambandha*' by *ṣaṣṭhī-tatpuruṣa-samāsa.*

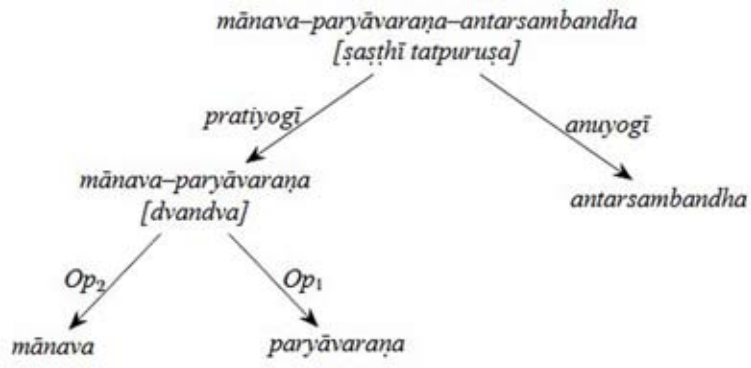

Fig. 4. The compound structure in (Ex. 13)

Let us see USR in Table 7:

Table 7. Universal Semantic Representation of the (Ex. 13)

<sent_id= Geo_nios_1ch_0056c>							
Cons	**Index**	**Semcat**	**MoSem**	**DepR**	**DisE**	**SpeakV**	**Cons**
anya_1	1	–	–	3:mod	–	–	–
mahatvapūrṇa_1	2	–	–	3:mod	–	–	–
vicāraka_1	3	anim	pl	13:k1	–	–	–
Bhūgola_1	4	–	–	13:k2	–	–	–
mānava_1	5	anim	–	–	–	–	**7:op1**
paryāvaraṇa_1	6	–	–	–	–	–	**7:op2**
[dvandva_1]	7	–	–	–	–	–	**9:pratiyogi**
antarsamban-dha_1	8	–	pl	–	–	–	**9:head**
[6-tat_1]	9	–	–	10:r6	–	–	–
rūpa_1	10	–	–	13:k7	–	–	–
paribhāṣita_1	11	–	–	–	–	–	13:kriyāmūla
kara_1-yā_hai_1	12	–	–	–	–	–	13:verbalizer
[cp_1]	13	–	–	0:main	–	–	–
</sent_id>							

As highlighted in Table 7, **[dvandva_1]** denotes a *dvandva-samāsa* between the mānava_1 and paryāvaraṇa_1. So, op1 and op2 in the

construction indicate the role of the mānava_1 and paryāvaraṇa_1 in the compound.

4.2.2. Samāsa-vṛtti-garbhaka anya-vṛtti (Compounds with other Constructions)

Samāsa-vṛtti-garbhaka anya-vṛtti refers to cases in which a compound first undergoes *samāsa-vṛtti* and subsequently serves as the base for another *vṛtti*. In such constructions, the compound meaning forms an intermediate stage in the derivation rather than the final interpretation. The complete meaning arises only after the application of the subsequent *vṛtti*, such as a derivational suffix. Thus, the semantic output of the compound functions as the input to a further grammatical operation.

• **Tatpuruṣa-garbhaka taddhita-vṛtti (Determinative compound with Derivational Suffixation)**

A *Tatpuruṣa-garbhaka taddhita-vṛtti* denotes a multi-layered grammatical process where a determinative compound, such as *karmadhāraya-samāsa*, is first formed to create a unified base, to which a *taddhita* (secondary derivational) suffix is subsequently appended to derive a new lexeme. For Instance, let us see in (Ex. 14).

(Ex. 14) *uṣṇa kaṭibandhīya cakravāta śītoṣṇa kaṭibandhīya cakravāta se kaī bātoṃ meṃ bhinna hote haiṃ* [Geography (316)].

Hin: *uṣṇa kaṭibandhīya cakravāta śītoṣṇa kaṭibandhīya cakravāta se kaī bātoṃ meṃ bhinna hote haiṃ.*

Gls: tropical *adj.* equatorial adj. cyclone *nom. m. pl.* temperate *adj.* equatorial *adj.* cyclone *nom. m. pl.* several *adj.* aspects *loc. f. pl.* different *adj.* be *pres. 3 pl.*

Eng: Tropical cyclones are different from temperate cyclones in several aspects.

In (Ex. 14), the word *kaṭibandhīya* is derived through *taddhita-vṛtti*. Here, the *taddhita* suffix *īya* (*cha*) is added to the base *kaṭibandha* in the sense of "*tatra bhavaḥ*" ("existing there"). This semantic interpretation is rooted in the Pāṇinian grammatical tradition, where *taddhita* suffixes are prescribed for specific semantic relations (Aṣṭādhyāyī 4.3.53). Accordingly, *kaṭibandhīya* denotes something associated with or existing in relation to a belt (*kaṭibandha*). That refers to *cakravāta* (cyclone), which exists in the *kaṭibandha*. *Uṣṇa* (hot) is a modifier of the *kaṭibandha.* But, if the *kaṭibandha* has an

external (outside of *taddhita-vṛtti*) modifier '*uṣṇa*', *the taddhita* suffix '*-īya*'*(cha)* cannot be added to '*kaṭibandha*' due to Rule 1.a.

Suppose one formed the word '*kaṭibandhīya*' *by* adding the *taddhita* suffix '*-īya*'*(cha)* without considering the modifier *uṣṇa*. Now, if one wishes to connect an external modifier *uṣṇa* (hot) with *kaṭibandha* in *kaṭibandhīya*, still *uṣṇa* can not be connected due to Rule 1.b. Therefore, first of all, it is necessary to form the compound word '*uṣṇa-kaṭibandha*' using *karmadhāraya samāsa* between the words *uṣṇa* and *kaṭibandha*. After that, the word '*uṣṇa-kaṭibandhīya*' will be meaningfully derived by adding the *taddhita* suffix '*-īya*'*(cha)* to '*uṣṇa-kaṭibandha*'. To clarify this idea, consider the following Figure 6.

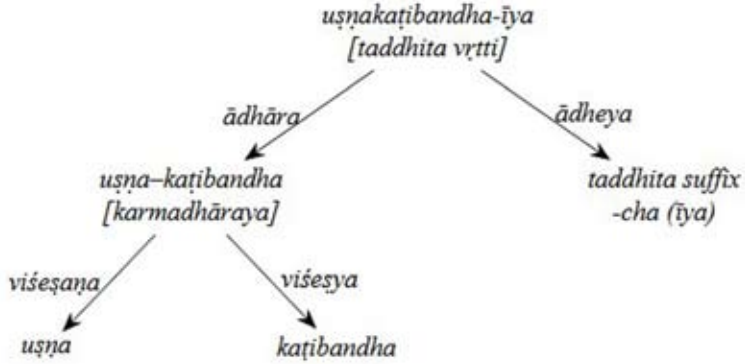

Fig. 5. The *vṛtti* structure in (Ex. 14)

Let's see the USR in Table 8.

Table 8. Universal Semantic Representation of the (Ex. 14)

<sent_id= Geo_nios_11ch_0295>							
Conc	**Index**	**SemCat**	**MoSem**	**DepR**	**DisE**	**SpeakV**	**Cons**
uṣṇa_1	1	–	–	–	–	–	**3:mod**
kaṭibandha_1	2	–	–		–	–	**3:head**
[karmadhā-raya_1]	3	–	–	–	–	–	**4:location**
[taddhita_1]	4	–	**tatra-bhavaḥ**	5:mod	–	–	–
cakravāta_1	5	–	–	14:k1	–	–	–
śītoṣṇa_1	6	–	–	–	–	–	**8:mod**

kaṭibandha_1	7	–	–	–	–	–	**8:head**
[karmadhā-raya_2]	8	–	–	–	–	–	**9:location**
[taddhita_2]	9	–	**tatra-bhavaḥ**	10:mod	–	–	–
cakravāta_1	10	–	–	13:k5	–	–	–
kaī_1	11	–	–	12:quant	–	–	–
bāta_1	12	–	–	14:k7	–	–	–
bhinna_1	13	–	–	14:k1s	–	–	–
ho_1-tā_hai_1	14	–	–	0:main	–	–	–
</sent_id>							

As highlighted in Table 8, **[karmadhāraya_1]** denotes a *karmadhāraya-samāsa* between the words *uṣṇa* and *kaṭibandha*. In the construction label, **mod** and **head** relations are assigned for *uṣṇa* and *kaṭibandha*, respectively, with reference to Index **3** (karmadhāraya). **[taddhita_1]** denotes a *taddhita-vṛtti* of [karmadhāraya_1] (i. e., *uṣṇa-kaṭibandha*) where *taddhita* suffix '*-īya*'*(cha)* is to be added. Here, *uṣṇa-kaṭibandhīya* (belonging to the hot belt / torrid zone) is derived through a *taddhita suffix*. The derivation encodes the semantic relation *tatra-bhavaḥ* (that which exists in a particular place), indicating association with the *uṣṇa-kaṭibandha* (hot or equatorial belt). Accordingly, this relation is assigned to **[taddhita_1]** (*uṣṇa-kaṭibandhīya*) in the **Morpho-Semantic** annotation.

The USR representations presented above serve as input to the natural language generation (NLG) system. In the following section, these representations are used to generate output sentences using a large language model (Gemini 2.5). The generated outputs are compared with reference translations and outputs from standard systems, such as Google Translate and GPT-based models, to evaluate the effectiveness of the USR-to-NLG pipeline, particularly in preserving compound structures and their internal relations.

4.3. From USR to Multilingual Generation

The Universal Semantic Representation (USR) serves not only as a framework for semantic analysis but also as an intermediate representation for multilingual natural language generation. Since USR encodes language-independent semantic relations, it enables the

generation of equivalent sentences across languages while preserving the underlying meaning, including compound structures and their internal relations.

The process consists of three stages: sentence input, USR construction, and natural language generation (NLG). In the first stage, a language-specific sentence is taken as input. In the second stage, the sentence is converted to USR manually or using an automatic USR builder tool, in which semantic relations, compound structures, and nested dependencies are represented in a language-independent format. In the final stage, the USR is provided as structured input to a large language model (LLM), specifically Gemini 2.5, for natural language generation. The USR representation is supplied in a formatted schema (including nodes, relations, and compound markers), enabling the model to generate output sentences that preserve both semantic content and structural relations. During NLG, compound-related features encoded in USR guide surface realisation in the target language. Based on the internal structure, such as modifier-head relations and nested vṛttis, the system determines whether the output should be realised as a compound form or as an analytic expression. This ensures that complex nominal structures are generated without loss of internal semantic relations, which are often simplified in standard translation systems.

The following examples illustrate the outputs generated through the USR-to-NLG pipeline using Gemini 2.5. For each case, the USR representation serves as structured input, and the resulting sentences are analysed with respect to the preservation of compound structures and their internal relations. The generated outputs are first examined to evaluate how effectively the system realises the encoded semantic and structural information. For example, consider Example 10, 13 and its corresponding USR (Table 3) and (Table 6).

Input: Table 3. Universal Semantic Representation of Example 10.

Output (USR to NLG – English):

Early scholars were experts of descriptive geography.

Output (USR to NLG – Hindi):

prāraṃbhika vidvāna varṇanātmaka bhūgola ke vettā the

The generated outputs correctly realise the compound structure "*varṇanātmaka bhūgola ke vettā*", preserving the hierarchical relation between "*varṇanātmaka bhūgola*" (descriptive geography) and "*vettā*" (experts). In English, this relation is expressed through an analytic construction ("experts of descriptive geography"), while in Hindi, the

compound structure is preserved in its natural form. The genitive marker "***ke***" and "***of***" are appropriately generated in Hindi and English, respectively, reflecting the underlying relational structure encoded in the USR representation.

Input:

Table 6. Universal Semantic Representation of Example 13.

Output (USR to NLG – English):

Other important thinkers have defined geography as the form of human-environment interrelations.

Output (USR to NLG – Hindi):

Anyā mahatvapūrṇa vicārakoṃ ne bhūgola ko mānava aura paryāvaraṇa ke antarsambandhoṃ ke rūpa meṃ paribhāṣita kiyā hai

The generated outputs correctly realise the compound structure "***mānava-paryāvaraṇa antarsambandha***", preserving the coordinated (dvandva) relation between "***mānava***" and "***paryāvaraṇa***". In Hindi, the dvandva relation is explicitly realised through the conjunction "***aura***", while the *ṣaṣṭhī* relation is expressed by the genitive marker "***ke***", resulting in the expression "***mānava aura paryāvaraṇa ke antarsambandha***". In English, this structure is represented through the analytic expression "human-environment interrelations," where coordination is preserved through hyphenation and the relational meaning is maintained. These realisations demonstrate that the USR-to-NLG process accurately maps compound relations into appropriate surface markers across languages. These examples further demonstrate that the USR-based approach effectively preserves both coordination and relational structure in compound expressions, ensuring accurate and natural realisation across languages.

4.4 Distribution of Samāsa-vṛtti in NIOS Data

NIOS data has been considered in this paper to analyse compounds. This study examines the occurrence and distribution of compound words in the NIOS Geography text in Hindi, focusing on four selected chapters. The dataset comprises 1504 sentences, among which 463 compound words appear in various forms, such as *tatpuruṣa, karmadhāraya, madhyama-pada-lopi*[11], *dvandva, upa-*

[11] *Madhyama-pada-lopi-samāsa* refers to a type of compound formation in which two words first combine to form a compound word. This compounded form subsequently enters into a *karmadhāraya* compound with another word. In the resulting *karmadhāraya-samāsa*, the medial component, which was the final (second) component in the first compound, is elided. A

pada[12]*, bahuvrīhi* and *kevala samāsa*. The analysis considers only separated or hyphenated compounds, excluding single-word compounds inherently fused in writing. Let us see the major compounds through Figure 6:

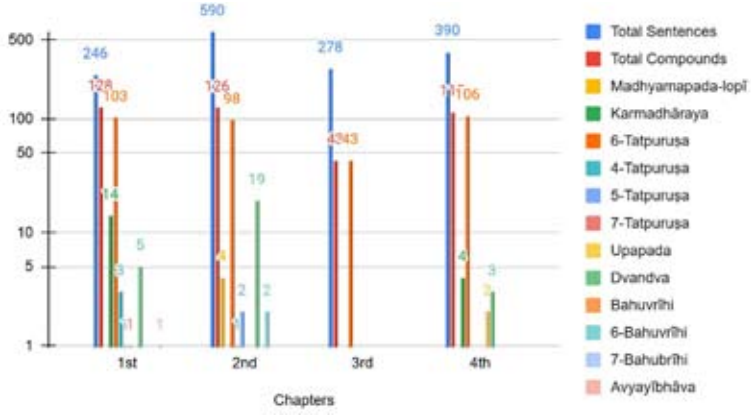

Fig. 6. Total number of major compounds and their types in the NIOS text Geography in Hindi

compound formed through such deletion of the medial element is known as *madhyama-pada-lopi-samāsa*. In the first stage of formation, two lexical items combine to form a compound, and this compound may belong to any *samāsa* type. In the second stage, this already compounded form combines with another word by a *karmadhāraya-samāsa*. It is during this second stage of compounding that the medial constituent is deleted: *śākapriyaḥ pārthivaḥ* → *śākapārthivaḥ* [Dīkṣita et al. 2022]. Here, the word *śāka-priya* is itself a *bahuvrīhi* compound, analysable as *śākaṃ priyaṃ yasmai* (one for whom vegetables are favourite). Subsequently, when *śāka-priya* and *pārthiva* combine through a karmadhāraya compound, the medial element *priya* is elided, resulting in the form *śāka-pārthiva*.

[12] **Upapada Samāsa** is a specific category of tatpuruṣa compound defined by the *Pāṇinian* rule **"Upapadamatiṅ" (P. 2.2.19)**, where a subanta (inflected noun) acts as a nerest dependent (upapada) compound with a verbal derivative (*kṛdanta*). For example, in **Kumbhakāraḥ** (*Kumbhaṃ karoti iti*; who makes a pot), meaning Potter, the suffix aṆ is added to the root kṛ only because of the preceding object, Kumbha. Similarly, in **Jaladaḥ** (*Jalaṃ dadāti iti*; that who gives water), meaning Cloud, the word *Jala* (water) serves as the *upapada* for the root *dā* (to give), illustrating how these compounds inherently encode *kāraka* relations, such as *karma* (object), within a single lexical unit.

Here is the Summary of **Figure 9:**

Chapter 1st: This chapter contains 246 sentences and 128 compounds, resulting in a high compound ratio of 52 %. The substantial presence of *Tatpuruṣa* and *Madhyamapadalopin* compounds indicates a rich syntactic structure.

Chapter 2nd: This chapter has 590 sentences, the highest among all, with 126 compounds, leading to a compound ratio of 21 %. The lower compound density suggests simpler syntactic constructions despite the large text volume.

Chapter 3rd: It comprises 278 sentences but with only 43 compounds, resulting in the lowest compound ratio of 15 %. The limited occurrence of compounds reflects a relatively straightforward and less compact linguistic style.

Chapter 4th: This chapter comprises 390 sentences and features 106 compounds, resulting in a compound ratio of 27 %. The occurrence of *Tatpuruṣa*, *Dvandva*, and *Upapada* compounds indicates moderate syntactic complexity.

Within the compounds mentioned in **Figure 6**, certain words have nested compounds according to the lexical meaning and as per Rule No. 1. These compounds are explained with examples in Section 4.2. Now, see the types and frequency of nested compounds in **Figure 7**.

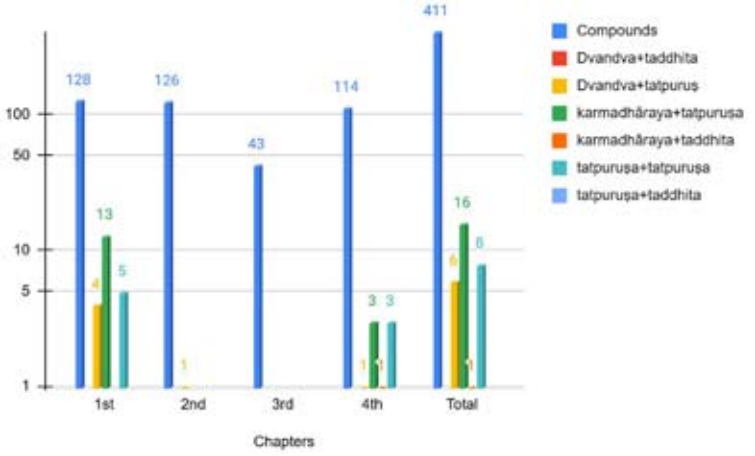

Fig. 7. Total number of Nested Compounds and their types in the NIOS text Geography in Hindi

The 1st chapter has the highest compound diversity, featuring *tatpuruṣa-garbhaka-tatpuruṣa* (13 examples) and *karmadhāraya-garbhaka-tatpuruṣa* (5 examples), indicating complex syntactic structuring. The 4th chapter also shows moderate richness with *tatpuruṣa-garbhaka-tatpuruṣa* (3 examples) and *karmadhāraya-garbhaka-tatpuruṣa* (3 examples). In contrast, the 2nd and 3rd chapters display minimal diversity, each containing only one instance of a complex compound type, suggesting simpler syntactic patterns. Overall, the 1st and 4th chapters are linguistically dense, while the 2nd and 3rd lean towards straightforward sentence structures.

5. Conclusion

This study has demonstrated that compound formation, particularly nested *vṛtti* structures, can be systematically analysed and represented within the Universal Semantic Representation (USR) framework by incorporating principles from the Pāṇinian grammatical tradition. The primary purpose of grammar, as a structured system, is to ensure clarity, precision, and consistency in the organisation and interpretation of language. In this regard, Pāṇini's grammar provides a highly systematic and rule-governed framework for both syntactic and semantic analysis, which proves particularly effective for modelling compound structures. By identifying specific types of nested compounds and formulating constraints, such as Rule 1 and its exceptions, this study establishes a principled method for capturing internal hierarchical relations in compounds. In particular, the organisation of *samāsa* based on the principle of *sāmarthya* and the application of Rule 1, wherever applicable, provide a structured mechanism for compound analysis. At the same time, the selective relaxation of this rule in cases such as *nitya samāsa* and relational constructions (e.g., *grāmasya prati-vṛkṣam*, *devadattasya guru-putraḥ*) highlights the flexibility required to account for linguistic realities. The analysis further shows that such a rule-based approach enables an accurate representation of both structural and semantic dependencies in USR. An evaluation using the USR-to-NLG pipeline with a large language model demonstrates that explicitly encoded compound relations significantly improve the quality of generated outputs. Unlike standard translation systems, which tend to simplify internal structures, the proposed approach preserves compound integrity and produces semantically faithful outputs across languages.

Overall, this study highlights the importance of integrating traditional grammatical insights into computational frameworks. Analysing compounds in USR under systematically defined conditions not only simplifies complex linguistic structures but also enhances their usability in natural language processing and machine learning applications. Future work may extend this approach to broader datasets and explore automated methods for detecting and encoding nested compound structures in multilingual contexts.

6. Appendix

The terms used in the USR are specific to the domain and may require clarification for readers unfamiliar with the context. Let us see the complete definition and explanation provided in the Appendix below, which will help clarify their meaning and relevance within this study. Look at Table 9.

Table 9. Definition of the specific terms and Tags used in the USR

Tag	Relation/Full form	Definition
anim	Animacy	living beings unless it is a proper noun.
pl	Plural	–
k7t	*kāla-adhikaraṇa-kāraka*	Time of the event.
quant	Quantifier	A limiting noun modifier expresses quantity.
k4a	*anubhava-kartā*	Experiencer
k2	*karma-kāraka*	That which the agent most seeks to encompass with his action is called *karma-kāraka* ‘object’.
rkl	*kālala-kṣaṇa*	Time is the referent of actual temporal information of the event.
op1	First opponent	One of the components of the conjunctive/ disjunctive unit.
op2	Second opponent	One of the components of the conjunctive/ disjunctive unit.
mod	*viśeṣaṇa*	The modifier.
head	*viśeṣya*	The modified.

0:main	*mukhya viśeṣya*	The main head in the sentence
k1s	*kartā-samāna-adhikaraṇa*	The kartā and its viśeṣaṇa reside in the same locus when the verb is copulative.
jñeya	Knowable	The object of the knowledge.
jñātā	Knower	A person who knows.
k1	*kartā-kāraka* (agent)	The most independent participant in an action.
k2	*karma-kāraka* (object)	locus of the result of the action.
rt	*tādarthya*	Purpose of the event.
k7	*viṣaya-adhikaraṇa-kāraka*	Location elsewhere.
op1	–	First of the components of the conjunctive/disjunctive unit.
op2	–	The second component of the conjunctive/disjunctive unit.
yoc	–	Years of the century.
per	–	Person
male	–	Masculine
[6-tat_1]	–	*ṣaṣṭhī-tatpuruṣa-samāsa* (where the non-head element bears implicit genitive case)
prayojya	–	The object/action for a purpose.
prayojana	–	The purpose of some object/action.
upakārya	–	The one being helped.
upakāraka	–	The one providing help.
pratiyogi	–	The entity whose relation is specified or intended is called the Pratiyogi of that relation.
ādhāra	–	Container / location, where something exists.
ādheya	–	Contained / located, which exists somewhere.

REFERENCES

Ācārya D. (ed.) (2014), *Nyāya-darśanam (Vātsyāyana-bhāṣya-sahitam)*, 7th ed., Chowkhamba Sanskrit Bhavan, Varanasi.

Amarasiṃha (1971), *Amarakosha*, Ed. by A. A. Ramanathan, Madras: The Adyar Library and Research Centre, available at: https://archive.org/details/amarkosh_english_202311/mode/2up (accessed 13.05.2026).

Dīkṣita B., Sarasvatī J., Dīkṣita V. and Bhaṭṭa N. (2022), *Vaiyākaraṇa-siddhānta-kaumudī with Tattva-bodhinī, Bāla-manoramā and Laghu-śabdendu-śekhara* (G. Śāstrī, S. Śāstrī and B. Śāstrī, eds), reprint, Vol. 1, Chaukhamba Surabharati Prakashan, Varanasi.

Bhaṭṭa N. (2006), *Parama-laghu-mañjūṣā* (Mishra Vanshidhar, ed.), First ed., Chaukhamba Sanskrit Pratishthan, Delhi.

Bhattāchārya G. and Mishra J. (2016), *Vyutpatti-vāda jayākhyā-vyākhyayā sahitā* (U. Mishra, ed.), reprint, Vol. 1, Vāṇī Vilāsa Prakāśana, Varanasi.

Bhikshu G. (1983), *Sarvatantra-siddhānta-padārtha-lakṣaṇa-saṅgraha*, 2nd ed., Gujarati Patra, Mumbai.

Cardona G. (1997), *Pāṇini: His Work and Its Traditions. Volume One: Background and Introduction*, Motilal Banarsidass, Delhi.

Garg K., Paul S., Sukhada, Bawahir F. and Kumari R. (2023), "Evaluation of Universal Semantic Representation (USR)", *ACL Anthology*, available at: https://aclanthology.org/2023.dmr-1.2/ (accessed 3 February 2026).

Geography (316) Syllabus (n.d.), *The National Institute of Open Schooling (NIOS)*, available at: https://nios.ac.in/online-course-material/sr-secondary-courses/Geography-(316).aspx (accessed 3 February 2026).

Hock H. H. (1991), *Principles of Historical Linguistics*, 2nd ed., Mouton de Gruyter, Berlin.

Iyer K. A. S. (1974), *The Vākyapadīya of Bhartṛhari, Chapter III (English Translation)*, Motilal Banarsidass, Delhi.

Jha V. N. (ed. and transl.) (2010), *Tarkasaṅgraha of Annambhatta: English Translation with Notes*, Chinmaya International Foundation Shodha Sansthan, Ernakulam, Kerala.

Kulkarni A. (2007), "Sanskrit and Computational Linguistics", paper presented at the First International Symposium on Sanskrit Computational Linguistics, Paris, 30 October 2007.

Lowe J. J. (2015), "The Syntax of Sanskrit Compounds", *Language*, Vol. 91, No. 3, pp. e71–e115, available at: https://doi.org/10.1353/lan.2015.0034 (accessed 4 June 2026).

Miśra R. (2015), *Aṣṭādhyāyī-sūtra-pāṭhaḥ (Pāṇini-muni-praṇītaḥ) Vārtika-gaṇapāṭha-dhātupāṭha-liṅgānuśāsana-vimarśa-sahitaḥ*, Motīlāla Banārasīdāsa, New Delhi.

Molina-Muñoz P. (2013), "Sanskrit Compounds and the Architecture of the Grammar", in *Grammatica and Verba: Glamour and Verve: Studies in South Asian, Historical, and Indo-European Linguistics in Honour of Hans Henrich Hock on the Occasion of His Seventy-Fifth Birthday* (S. Fukuda Chen and B. Slade, eds), pp. 181–201, Beech Stave Press, Ann Arbour.

Pañcholi B. (ed.) (2011), *Vaiyākaraṇa-bhūṣaṇa-sāraḥ with Prabhā Commentary of Bālakṛṣna-pañcolī and Darpaṇa of Harivallabha-śāstrī*, Chowkhamba Sanskrit Sansthan, Varanasi.

Patañjali, Kaiyaṭa and Bhaṭṭa N. (2018), *Vyākaraṇa-mahābhāṣyam, Bhāṣya-pradīpa-Pradīpodyota-sahitam* (B. S. Jośī, ed.), reprint, Vol. 2, Caukhambā Saṃskṛta Pratiṣṭhāna, Delhi.

Sastri P. S. S. (2015), *Lectures on Patañjali's Mahābhāṣya*, Vol. 5 (Āhnikas 15–22), The Kuppuswami Sastri Research Institute, Chennai.

Staal F. (1972), "The Science of Language", *Scientific American*, Vol. 227, No. 3, pp. 73–82.

Sukhada (2017), *A Paṇinian Perspective to Information Dynamics in Language: Mapping Structures between English and Hindi*, PhD dissertation, International Institute of Information Technology (IIIT), Hyderabad.

Varadarājācārya (2004), *Laghu-siddhānta-kaumudī* (R. D. Paṇḍeya, ed.), 33rd ed., Gita Press, Gorakhpur.

Viśvanātha Pañcanana and Kṛṣṇavallabhācārya (2018), *Nyāya-siddhānta-muktā-valī with the Commentary Kiraṇā-valī* (N. Śāstrī and Ś. Śāstrī, eds.), reprint, Vol. 1, Caukhambā Saṃskṛta Saṃsthāna, Varanasi.

Yāska (2004), *Nirukta-śāstram* (Bhagavaddatta, ed.), Second ed., Vol. 1, Ramlal Kapoor Trust, Delhi.

Website

Ashtadhyayi.com (n.d.), *Panini's Ashtādhyāyī with Annotations and Tools*, available at: https://ashtadhyayi.com (accessed 3 February 2026).

Сударшан Ґаутам, Сукхада

ФОРМУВАННЯ ТА РЕПРЕЗЕНТАЦІЯ СКЛАДНИХ СЛІВ ДЛЯ КОМП'ЮТЕРНОЇ ОБРОБКИ: ПІДХІД УНІВЕРСАЛЬНОГО СЕМАНТИЧНОГО ПРЕДСТАВЛЕННЯ З ПАНІНІЙСЬКОЇ ПЕРСПЕКТИВИ

Розуміння семантичної структури складних утворень є центральною проблемою як у теоретичній, так і в комп'ютерній лінгвістиці. Universal Semantic Representation (USR) забезпечує обчислювально придатну структуру для репрезентації значення у впорядкований та мовно-незалежний спосіб, спираючись на індійську граматичну традицію (Indian Grammatical Tradition, IGT). У цьому дослідженні принципи словоскладання Паніні застосовуються до структури USR з метою аналізу складних конструкцій з панінійської перспективи. Хоча такі принципи, як *ākāṅkṣā* (очікуваність) та *yogyatā* (семантична узгодженість), є фундаментальними для семантичної інтерпретації, дослідження особливо демонструє, як специфічні панінійські обмеження, включно з "*saviśeṣaṇānāṃ vṛttir na*" та "*vṛttasya vā viśeṣaṇayogo na*", допомагають розв'язувати неоднозначності сфери модифікатора в автоматизованих системах і семантичному парсингу. У статті також розглядається внутрішня семантична організація складних виразів та ієрархічні відношення, що виникають у вкладених і складних композитних конструкціях. Інтеграція цих положень у структуру USR демонструє, яким чином семантика складних утворень, відношення залежності та контекстуальна інтерпретація можуть бути систематично репрезентовані між мовами для комп'ютерної обробки. Приклади, взяті з підручників географії мовою гінді, ілюструють практичну застосовність запропонованого аналізу та підкреслюють значення панінійської граматики для розвитку комп'ютерних підходів до семантичної репрезентації та Natural Language Generation (NLG). У дослідженні проаналізовано 1504 речення з підручників географії мовою гінді та встановлено, що щільність складних утворень у технічних розділах сягає приблизно 52 %, що, своєю чергою, зумовлює необхідність систематичної логіки для репрезентації вкладених композитних структур і збереження семантичної цілісності. Результати свідчать, що інтеграція панінійських граматичних положень із USR робить вагомий внесок у моделі Natural Language Processing (NLP) для індійських мов шляхом підвищення семантичної точності, інтерпретаційної послідовності та комп'ютерної репрезентації складних мовних структур. Дослідження

підкреслює незмінну актуальність традиційних індійських граматичних теорій у сучасній комп'ютерній лінгвістиці та дослідженнях репрезентації знань.

Ключові слова: *vṛtti*, граматика Паніні, *ekārthībhāva-sāmarthya*, словоскладання, семантична сумісність, Universal Semantic Representation

Стаття надійшла до редакції 04.02.2026