OntoSenseNet: A Verb-Centric Ontological Resource for Indian Languages

by

Jyoti Jha, Sreekavitha Parupalli, Radhika Mamidi

in

19th International Conference on Computational Linguistics and Intelligent Text Processing
(*CICLing-2018*)

Hanoi, Vietnam

Report No: IIIT/TR/2018/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2018

OntoSenseNet: A Verb-Centric Ontological Resource for Indian Languages

Jyoti Jha*, Sreekavitha Parupalli**, Navjyoti Singh

Center for Exact Humanities
IIIT-Hyderabad
jyoti.jha@research.iiit.ac.in, sreekavitha.parupalli@research.iiit.ac.in,
navjyoti@iiit.ac.in

Abstract. Following approaches for understanding lexical meaning developed by Yāska, Patanjali and Bhartrihari from Indian linguistic traditions and extending approaches developed by Leibniz and Brentano in the modern times, a framework of formal ontology of language was developed. This framework proposes that meaning of words are *in-formed* by intrinsic and extrinsic ontological structures. The paper aims to capture such intrinsic and extrinsic meanings of words for two major Indian languages, namely, Hindi and Telugu. Parts-of-speech have been rendered into sense-types and sense-classes. Using them we have developed a gold-standard annotated lexical resource to support semantic understanding of a language. The resource has collection of Hindi and Telugu lexicons, which has been manually annotated by native speakers of the languages following our annotation guidelines. Further, the resource was utilised to derive adverbial sense-class distribution of verbs and kāraka-verb sense-type distribution. Different corpora (news, novels) were compared using verb sense-types distribution. Word Embedding was used as an aid for the enrichment of the resource. This is a work in progress that aims at lexical coverage of language extensively.

1 Introduction

The concept of 'meaning' has been discussed for a long time. Cognitively it can be understood to have an *intensional* or *extensional* form. Frege [1] discussed the idea of sense and reference. He called 'sense' as *intensional* meaning and 'reference' as *extensional* meaning. The meaning that has a constant value in an expression is *intensional*, whereas the meaning that is contributed by the real world to the mental concept is *extensional*. Two words are said to be extensionally equvivalent if they refer to the same set of objects, whereas if they share the same features then they are intensionally equivalent. According to Frege every signficant linguistic expression has both 'sense' and 'reference'. The other

* author has contributed to resource development for Hindi
** author has contributed to resource development for Telugu

theories of meaning are correspondence theory, consensus theory, constructivist theory etc. All these account for extensional meaning.

Meaning of a word in a language is generally derived from dictionary or from a context it is used in. Speaking from an ontological viewpoint, the meaning of a word can be understood based on its participation in classes, events and relations. In order to manipulate language computationally at the level of lexical meanings, Otra [2] developed Formal Ontology of Language. It considers meaning to have an intrinsic form. According to the theory proposed, meanings have primitive ontological forms. It is language independent and aims at extensive coverage of language.

This paper uses the idea of Formal Ontology of Language to develop lexical resource for Hindi and Telugu. Section 2 discusses the previous works that have been done in order to specify meaning of a word and development of various lexical resource. Section 3 of the paper discusses the Formal Ontology of Language, as proposed by Otra [2]. Section 4 talks about data acquisition for Hindi and Telugu and it shows how sense identification is done for different parts-of-speech. kāraka information for sense-types of verbs have been extracted from Hindi corpus. OntoSenseNet, a user interface has been built for our ontological resource. Section 5 shows validation of the resource based on inter-coder agreement. Section 6 demonstrates enrichment of the resource using word embeddings. Representation of verbs through their adverbial class distribution has been studied. Different corpora(news, novels) have been compared using frequency profiling of verb sense-types. Section 7 outlines conclusion and future work. IAST based transliteration[1] for Devanagari and Telugu scripts has been used in the paper.

2 Previous Work

In this section we discuss several attempts that have been made by various researchers in order to specify meaning of a word. Considering verb as the core of a language, some linguists derived different classifications. Along with different kinds of verb classifications, there have been approaches to derive semantic primitives.

Levin's classification Levin assumed that syntactic behavior of a verb is semantically determined [3]. He classified meanings of about 3,000 English verbs. They were composed into 50 primary-classes and 192 sub-classes using methodical study of 79 diathesis alternations. It mainly deals with verb taking noun and prepositional phrase complements. Although it can empirically classify verbs but it only captures some facets of semantics [4]. It does not include verbs taking ADJP, ADVP, predicative, control and sentential complements and is highly language dependent.

VerbNet classfication VerbNet [5] is a hierarchical verb lexicon that represents verbs syntactic and semantic information. In each verb class, thematic roles are used to link syntactic alternations to semantic predicates. However,

[1] `https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration`

it contains limited coverage of lemmas and for each lemma the coverage of the senses are limited. Since it has been inspired from Levin's verb classification it is also language dependent.

WordNet It is a lexical database inspired by psycholinguistic theory of human lexical memory [6]. Words are organized as synsets. These synsets represent lexicalised concepts which are organized into synonym sets. These synsets are connected to other synsets by means of semantic relations. Nouns are organized as hypernymy and hyponymy relations. Verbs are organized as hypernym, troponym, entailment and coordinate terms. It does not have classification of adverbs. It also lacks information about verb syntax and is also language specific.

Wierzbicka's semantic primitives The concepts that can be innately understood without any further decomposition are Semantic Primes. The widely used example to explicate this concept is the verb '*touching*'. Its meaning can be readily understood, however a dictionary might define '*touch*' as "to make contact" and '*contact*' as "touching", provides no information if neither of these words are understood. The theory of semantic primes was introduced by Wierzbicka, [7]. It has been criticised for its reductive approach and is limited by its generative coverage in any language [8].

The limitations of the above theories in terms of being language specific and limited coverage in a language led to the formulation of Formal Ontology of Language by Otra [2]. In the proposed theory, the meaning are considered to have primitive ontological forms and they are independent of a language. Each of the parts-of-speech are organised as *types* or *classes*. To derive the intensional meaning of a sentence, one has to consider the relation between different parts-of-speech e.g the relation between verb-adverb, noun-adjective, verb-noun. The relations between the points are also considered to have ontological forms, which can help specifying meaning at a sentence level. The next section discusses the theory of Formal Ontology of Language, as introduced by Otra.

3 Formal Ontology of Language

In a language one can describe a state of affairs using different verbs, hence there is a verbal ambiguity. To derive the universal verb there have been several discussions in Greek and Indic traditions. While in Greek tradition 'be' is considered as the universal verb [9], on the other hand Indic tradition considers 'happening (bhavati)' as the universal verb. Let us consider a question "what are you doing?". This can be answered with the verb 'do'. Hence one can say that 'do' can be a primitive sense that will be present in every verb. However, Patanjali mentions three verbs that cannot be the answer to the above question. These are (1) being/existence (asti), (2) presence (vidyate), (3)happening (bhaāva). According to Bhartrihari, verb has sense of sequence and state. Hence, it has a sense of happening making it the universal verb. Linguistic traditions in India have long regarded verb as the centre of language (in both syntactic and semantic terms) right from Yāska , Pānini, Patanjali and Bhartrihari [10], [11], [12]. Meaning

of verbal element is seen as *bhāva* (happening) as opposed to *sattā* (being) which stands behind nominal elements [13], [14]. Later, independent of linguistic discourse, logicians brought out hundreds of *bhāva-s* (happening) with atomic transformational structure <*entity*$_1$|*entity*$_2$>like cause|effect, part|whole, predecessor|successor, qualifier|qualified, ascribed|ascriber, locus|located, etc. These are atomic discriminants which form elementary meanings. Meanings are not seen as an entity like semantic primitive [7] but as a unified discriminative structure with a form <*entity*$_1$|*contiguous with*|*entity*$_2$, in context of *continuum*>. For example, verb 'move' has a sense <predecessor state|*contiguous with*|successor state, in context of 'move' *continuum*>. Its meaning is a continuant feel of motion punctuated by discriminating logical structure of predecessor and successor states. One can read in its meaning such discrimination points. Leibniz called such punctuations as actual points [15] as endeavours, as ontologically vacuous, as different from Euclidean points. Brentano [16] also built an idea of mental continuants as punctuated with *modo recto* and *modo obliquo*. These boundaries, punctuations or points are ontological as they vacuously discriminate ontic entities or states which are felt as continuous. Otra [2] built formal ontology of lexical meaning using such punctuational boundaries.

Meaning of 'move' is always more than the discriminative senses we read in it. In the proposed formal ontology seven discriminant punctuations that are read in all verbs [2] are suggested and determined. When we say, 'rapidly move' or 'hesitantly move', we have done adverbial modification of the meaning of 'move' and have added new modifier|modified points in its meaning. When appending 'rapidly' we add temporal-feature-class in adverb whereas while appending 'hesitantly' we add force-feature-class in adverb to the meaning of 'move'. Even when we have discriminated temporal or force features, meanings of 'rapidly' and 'hesitantly' are more than their adverb sense-classes. Otra [2] has delineated four adverb classes of discriminant point. Verbs are also seen as contiguants of nouns in seven or eight case relations. These seven/eight classes of verb-noun pairing are further coincident boundaries in the meaning of articulation with the verb. Further, noun-noun pairs and noun-adjective pairs are more coincident points. Otra [2] also proposes twelve noun-verb types of sense-points in the ontology. The verb-centric formal ontology of meaning is based on sense-type and sense-class of punctuational boundaries that can be located in lexical meaning. TYPES and CLASSES are logical forms of intensional senses [[2], page 13,14,15]. Using the formal ontology we are building lexical resource of verbs, adverbs, nouns and adjectives in terms of basic discriminant points, which *in-form* their meaning. The formal ontology is language independent and thus we are developing the resource for several languages at once.

In the next section we discuss the resource creation for Indian languages, namely Hindi and Telugu using the proposed Formal Ontology of Language.

4 Resource Building

Otra [2] has developed the resource for English that has 3,867 verbs, 1,980 adverbs and 300 adjectives. In our resource, Sense-types of 3,152 Hindi and 3,379 Telugu verbs has been manually identified. Similarly manual identification of sense-classes of 2,214 Hindi and 101 Telugu adverbs has been done. Sense-types of 238 Hindi adjectives has been identified. Annotation for sense-types of Telugu adjectives is in progress. The annotators have native proficiencies in the corresponding languages.

4.1 Data Acquisition

Words were collected from different resources like dictionary, wordnet. These were further used to populate our resource. Since this is a work in progress, not all the words from the different resources have been added into our resource.

Hindi Distinct verbs, adverbs and adjectives for Hindi were collected from Hindi Wordnet [2] and Dictionary [3].

Telugu Telugu being a resource poor language, does not have a usable soft copy of Telugu-Telugu dictionary till date. We are developing it from the printed copy of "Sri Suryaraayandhra Telugu Nighantuvu" [17] for all of its eight volumes. Verbs, adverbs and adjectives have been completely populated in the usable soft copy from this dictionary, whereas work is still under progress for other parts-of-speech. Dictionaries for Telugu-Hindi, English-Telugu are available[4].

Table 1 shows the number of distinct verbs, adverbs and adjectives in each of the resources.

Table 1. Distinct number of verbs, adverbs and adjectives for Hindi, Telugu in different resources

Resource	Distinct Verbs	Distinct Adverbs	Distinct Adjectives
Hindi Wordnet	6778	2114	19190
Hindi-Shadsagara Dictionary	3529	1650	36398
OntoSenseNet(Hindi)	3152	2214	238
Telugu-Telugu Dictionary	8483	253	11305
Telugu Wordnet[5]	2795	442	5776
Telugu-Hindi	9939	142	1253
OntoSenseNet(Telugu)	3379	101	*Annotations are yet to be started*

[2] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[3] https://ia601603.us.archive.org/20/items/in.ernet.dli.2015.348711/2015.348711.Hindi-Shabdasagar.pdf

[4] https://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

[5] http://tdil-dc.in/indowordnet/

4.2 Sense-Identification

Verbs Different verbs can be used for describing the same situation. Thus verbs are colocative in nature. In a single verb many verbal sense points can be present and different verbs may share same verbal sense points. For example "walking", "running" entails a sense of 'move'. Verb like "studying" entails a sense of 'know' and 'do'. "Eating" entails a sense of 'do' and 'have'. Thus, verbs are organised as sense-types. Otra [2] has shown the existence of seven primitive sense-types of verbs. These seven sense-types of verbs have been derived by collecting the fundamental verbs used to define other verbs. These verbs were then grouped using intrinsic senses, which were designated to a particular sense-type. These sense-types are inspired from different schools of Indian philosophies. The seven sense-types of verbs are listed below with their primitive sense along with two Hindi and Telugu examples each.

1. Means|End - Do; khelanā (play), karanā (do); āduta (play), ceyuta (do)
2. Before|After - Move; bahanā (flow), calanā (walk); pāruta (flow), naduvuta (walk)
3. Know|Known - Know; jānanā (know), parakhanā (examine); ūhimcuta (imagine), parisīlimcuta (examine)
4. Locus|Located - Is; rahanā (stay), honā (happen); umduta (to be, stay), jaruguta (happen)
5. Part|Whole - Cut; kātanā (cut), mitānā (erase); koyuta (cut), vidipovuta (separate)
6. Wrap|Wrapped - Cover; jhāmpanā (cover), pahanānā (dress-up someone); mūyuta (cover) , ākramimcuta (contain forcefully)
7. Grip|Grasp - Have; pānā (get), lenā (take); bhayapaduta (fear), tīsukonu (take)

Each of the verbs can have all the seven dimensions of sense-types. The degree depends on the usage/popularity of a sense in a language. In our resource we have identified two sense-types of each verb, i.e. primary and secondary. Consider the verb 'dance' in the sentence "Madhuri is dancing gracefully". Here 'dance' involves a sense of movement which a doer does. Thus Before|After is a primary sense and Means|End is a secondary sense. For polysemous verbs, sense-type identification was done for each of their different meanings. For example, the verb "rap" has three meanings. Thus rap1, rap2, rap3 have been added in the resource along with its meaning sense-types.

1. rap1- Criticizing someone, Means|End and Know|Known.
2. rap2- To perform rap music, Means|End and Before|After.
3. rap3- To hit or say something suddenly and forcefully, Means|End and Part|Whole.

Sense-types for 3,152 Hindi and 3,379 Telugu verbs were manually identified. Figure 1 shows the sense-type distribution for English, Hindi and Telugu verbs in OntoSenseNet.

Adverbs Meaning of verbs can further be understood by adverbs, as they modify verbs. The sense-classes of adverbs are inspired from adverb classification

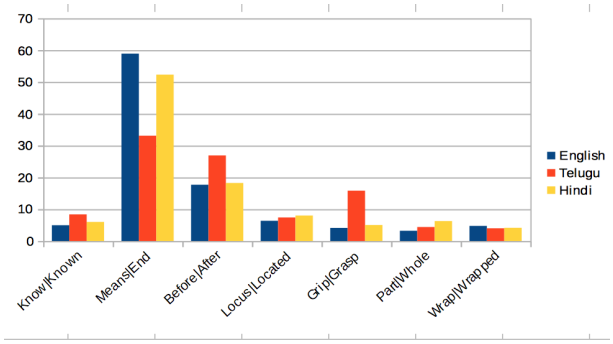

Fig. 1. Verb Sense-type distribution

in Sanskrit. Following are the identified sense-classes along with their fundamental sense, illustrated with English, Hindi and Telugu examples

1. Temporal - Adverbs that attributes to sense of time. e.g Never; sasamaya (timely); varusagā (continuously)
2. Spatial - Adverbs that attributes to physical space. e.g There; pās (near); davvu (far away)
3. Force - Adverbs that attributes to cause of happening e.g. Dearly; barbas (unwillingly); gattiga (tightly)
4. Measure - Adverbs dealing with comparison, judgement. e.g - Only,; lagbhag (approximately); gaddu (abundantly)

Sense-classes for 2,214 Hindi and 101 Telugu adverbs have been manually identified. Table 2 shows sense-class distribution of adverbs for English, Hindi and Telugu.

Table 2. Adverb Sense-Class Distribution

Sense-Class	English	Hindi	Telugu
Temporal	5.5	24.3	28.7
Spatial	2.7	13.5	12.8
Measure	39.4	32.2	31.6
Force	52.2	30	26.7

Sub-classfication of adverb sense-classes is being developed.

Kāraka Relation Kāraka are classes that are used for expressing relation between words in a sentence. Computational Paninian Grammar Framework describes kārakas as syntactico-semantic (or semantico-syntactic) relations between the verbs and their related constituents (generally nouns) in a sentence [18].

It describes eight types of kārakas:- k1: kartā (Nominative), k2: karma (Instrument), k3: karna (Ablative), k4:sampradāna (Possessive), k5:apādān (Objective), k6:sambandh (Dative), k7:adhikaran (Locative), k8:sambodhan (Vocative). Distribution of verb sense-types and kāraka were studied in Hindi novel corpora containing 3,39,057 words. This corpus was collected from Hindisamay[6] and was fully parsed using ISCNLP tagger[7]. For Telugu, development of treebank data and full dependency parser is still under process. Thus, Kāraka-Verb sense-types distribution of Telugu has not been extracted. Figure 2, shows verb sense-types distribution for different kāraka in Hindi. It shows that Locus|Located type of verbs have mostly occurred with a k1 relation with noun, whereas the k4 kāraka is hardly occurs in any verb-noun relation.

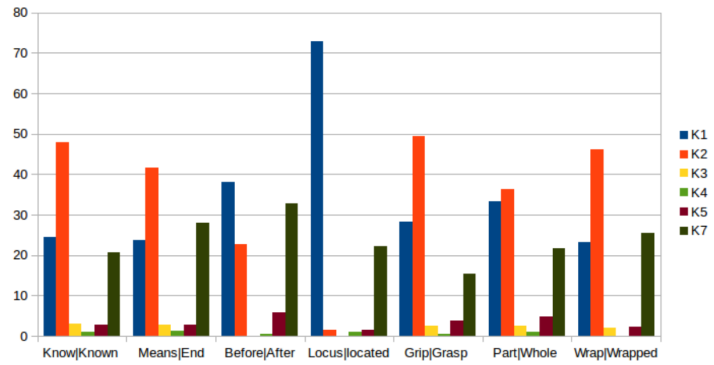

Fig. 2. Verb sense-type and Karaka Distribution

Adjectives Like verbs, adjectives are also colocative in nature. Otra [2] identifies 12 sense-types. However these can be reduced to 6 pairs. Following are the identified six sense-types pairs of adjectives along with their meanings and one English, Hindi and Telugu examples each.

- Locational - Adjectives that universalise or localise a noun e.g Local, doosrā (Other), nirdista (specific)
- Quantity - Adjectives that either qualify cardinal measure or quantify in ordinal-type e.g. - only, eka (one), okkati (One)
- Relational - Adjectives that qualify nouns in terms of dependence or dispersal e.g. similar, mātrihīn (without mother), vistrta (broad)
- Stress - Adjectives that intensify or emphasis a noun - e.g, strong, mazbūt (strong), gatti (strong)

[6] http://www.hindisamay.com/
[7] https://github.com/iscnlp/iscnlp

– Judgement - Adjectives that qualify evaluation or qualify valuation feature of a noun e.g - bad, acchā (good), mamci (good)
– Property - Adjectives that attribute a nature or qualitative domain of a noun. e.g - black, kālā (black), nallani (black)

Sense-types for 238 Hindi adjectives have been manually annotated. Sense-types for Telugu is yet to be started.

4.3 User Interface

A user interface, OntoSenseNet has been developed for this ontological resource. Input of a verb/adverb/adjective returns its corresponding sense-types/classes along with its illustrative meaning and example sentences. Further, development on crowdsourcing of the resource is being done where users can login and populate the data by providing examples to support their claims. This will be manually verified before adding to OntoSenseNet. For divided opinion, a discussion page would be provided.

5 Resource Validation

To show the reliability of the resource, Cohen's Kappa [19] was used to measure inter coder agreement. The annotation was done by one human expert and it was cross-checked by another annotator who was equally trained. Verbs and adverbs were randomly selected from our resource for the evaluation sample. The inter coder agreement for 500 Hindi verbs and 1,000 adverbs were 0.70 and 0.91 respectively. Similarly validation for 500 Telugu verbs was done, for which inter coder agreement was 0.82. Validation for both the language resources shows high agreement. [20]. Further validation of the resource is in progress.

6 Resource Enrichment and Utilization

6.1 Sense-Identification of Verbs and Adverbs using Word Embeddings

Word Embeddings have been widely used for extracting similar words [21]. Previous study has shown that word embedding has significant improvement over WordNet based measures [22]. We used this property to assign sense-type of verbs and sense-class of adverbs. This was done in order to facilitate the annotation task. However, this was further verified manually.

Method Hindi corpus was collected from Leipzig [8], Hindi wiki-dump [9] and Lindat [23]. It contains 3,73,45,049 sentences and 75,31,64,082 words. This corpus was fully parsed using iscnlp tagger[10]. Word2vec [24] was trained on this

[8] http://corpora2.informatik.uni-leipzig.de/download.html
[9] http://kperisetla.blogspot.in/2013/01/wikipediahi-offline-wikipedia-in-hindi.html
[10] https://github.com/iscnlp/iscnlp

corpus using CBOW technique and vector dimensions were 100.1,485 verbs and 1,054 adverbs were randomly chosen from the corpus for the sense-identification. Out of these, sense-type of 1,182 verbs and sense-class of 832 adverbs were already present in the resource. The sense identification for the remaining words were carried out.In order to identify sense of a word, its cosine similarity was calculated against the words whose sense were already present in the resource. Cosine similarity above 0.7 was considered. The maximum occurring sense in the similarity cluster was considered to be the potential sense of that word. This was subsequently verified manually. For example, sense-type of the Hindi verb 'cīranā' (tear) was not present in the resource(OntoSenseNet). The sense-type of those verbs were considered whose cosine similarity with 'cīranā' was above 0.7. The maximum occurring sense from these set of verbs was Part|Whole. Thus, the sense-type of 'cīranā' was assigned as Part|Whole. The above method was executed to identify sense-type of 303 verbs and sense-class of 222 adverbs. This method correctly identfied the sense -class of 220 adverbs and sense-type of 185 verbs. The sense identified for the words were finally incorporated in the resource. Table 3 summarises the statistics. Column A is part-of-speech. Column B shows number of words in that part-of-speech that were randomly sampled from the corpus. Column C shows number of words for which sense were already present in the resource. Column D shows number of words for which sense identification was carried out. Column E contains number of words whose sense were correctly identified by Word2Vec. Column F shows accuracy in percentage.

Table 3. Statistics for the sense-identification by Word2Vec

	A	**B**	**C**	**D**	**E**	**F**
Verb	1,485	1,182	303	185	61.056%	
Adverb	1,054	832	222	220	99.09%	

Table 4 and Table 5 shows the verb and adverb clusters, respectively. In each of the tables the similarity with the words in column-1 is above 0.7. Column 3 shows the maximum occurring sense-type/sense-class

Table 4. Similarity Cluster and the maximum occurring sense-type

Verb	**Verb-Clusters**	**Maximum occurring sense-type**
cīranā	nocanā, ghisanā, chedanā, khuracanā, pīsanā,phulānā	Part|Whole
jānanā	batānā, kahanā, lenā-denā, mālūma, mānanā, pūchanā	Know|Known

Table 5. Similarity Cluster and the maximum occurring sense-class

Adverb	Adverb-Clusters	Maximum occurring sense-class
tigunā	dogunā,dugunā,caugunā	Measure
yakāyaka	sahasā,ekāeka,acānaka	Temporal

6.2 Representation of verbs through adverbial semantics

Representation of verbs as a combination of their participatory adverb modifiers has not been exhaustively studied till now. Using our resource one can study the adverbial features of verb. We extracted Verb-Adverb relation from a fully parsed corpora.

Using full dependency parser of Hindi, 25,00,130 sentences were parsed. Verbs whose frequency was above 50 were considered in order to extract their modifying adverbs. The sense-class of these adverbs were then identified using our resource. Percentage of the frequency distribution of these sense-classes of adverbs for every verb was calculated. Table 6 shows few examples of represenation of verbs in terms of their frequency distribution of adverb sense-class.

Table 6. Percentage of frequency distribution of adverb sense-class of verbs

Verb	Temporal	Measure	Spatial	Force
cunanā	60.82	36.08	2.06	1.03
jānā	61.12	25.92	12.96	0
calanā	44.26	44.26	6.55	4.91
likhanā	32.07	50.31	10.69	6.91

Few examples of the verbs that were not modified by "Spatial" in the corpora are karwāne [to make someone do], chaunk [to be surprised], bachā [save], gher [circle] It is interesting to note that spatial and force were the only classes that did not modify some verbs. OntoSenseNet has verb-adverb pairing frequencies also.

Corpora Comparison Frequency distribution of verb sense-type and adverb sense-class across different types of corpora(novel, news) have been statistically studied. Previous works have shown the usage of frequency profiling [25] for comparing different corporas. We have used this approach to identify key ontological points that differ across corpora. Log-Likelihood estimation was calculated using contingency table for each of the verb sense-type. It was observed that Means|End sense-type is the most indicative of the news corpora that was collected using Hindi Treebank(3,65,431 words).

Table 7 shows the frequency distribution of sense-types of verbs and their log-likelihood.

Table 7. Frequency distribution and Log Likelihood

Sense-Type	**Novel**	**News**	Log-Likelihood
Means|End	25.215	38.270	+38523.04
Before|After	19.084	15.293	+9787.00
Part|Whole	5.917	5.736	+4290.64
Grip|Grasp	7.387	9.782	+9076.68
Locus|Locate	30.817	23.946	+14911.13
Know|Known	10.216	5.993	+2812.73
Wrap|Wrapped	1.360	0.977	+566.23

Furthermore, adverbial class distribution was observed in novels of two different authors. The adverbial distribution was considered for the most commonly occurring verbs in both the corpora. The difference in the use of adverbs may be accounted for different sociolinguistics aspects and can be applied in the study of social differentiation in the use of a language. Adjectival and kāraka information needs to be exploited further for a deeper insight into corpora comparison.Sense-identification for Telugu using Word2vec is in progress. Table 8 shows few examples of the adverbial sense-class distribution for the verbs used by both the authors.

Table 8. Adverbial sense-class distribution across different novels

Verb	**Adverb sense-class for Author-1**	**Adverb sense-class for Author-2**
letnā (To take rest)	Temporal	Temporal, Measure
khelnā (To play)	Temporal	Temporal, Measure, Force
likhnā (To write)	Measure	Temporal
girnā (To fall)	Temporal, Spatial, Force, Measure	Temporal, Force
jānā (To go)	Temporal	Temporal, Measure
denaā (To give)	Temporal, Measure	Temporal
sunanā (To listen)	Measure	Temporal, Measure

7 Conclusion and Future Work

In this paper we used Formal Ontology of Language to develop ontological resource for Hindi and Telugu. Logical forms of intensional senses were identified as type and class. The validation of this resource was done using Cohen's Kappa that showed higher agreement. The resource was used for extracting adverbial class distribution of verbs, kāraka-verb sense-type distribution from corpus. We

compared different corpora based on sense-type distribution of verbs. Novels of different authors were compared using sense-class of adverbs. The usage of different sense-class of adverbs across different authors indicates different sociolingustics aspect. However, this just covers a portion of a language. Adjectival and kāraka points will give a deeper insight. Further, validation and enrichment of the resource is in progress. Major work ahead is to find etymological, morphological and syntactic points. The resource can be utilized for word sense disambiguation, synonimity measure, cultural studies.

References

1. Gamut, L.: Logic, Language, and Meaning, volume 1: Introduction to Logic. Volume 1. University of Chicago Press (1991)
2. Otra, S.: Towards Building a Lexical Ontology Resource Based on Intrinsic Senses Of Words. PhD thesis, International Institute of Information Technology Hyderabad (2015)
3. Levin, B.: English verb classes and alternations. Volume 1. Chicago: University of Chicago Press (1993)
4. Korhonen, A., Briscoe, T.: Extended lexical-semantic classification of english verbs. In: Proceedings of the HLT-NAACL workshop on computational lexical semantics, Association for Computational Linguistics (2004) 38–45
5. Kipper, K., Dang, H.T., Palmer, M., et al.: Class-based construction of a verb lexicon. AAAI/IAAI **691** (2000) 696
6. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. International journal of lexicography **3** (1990) 235–244
7. Wierzbicka, A.: Semantic primitives. (1972)
8. Riemer, N.: Reductive paraphrase and meaning: A critique of wierzbickian semantics. Linguistics and philosophy **29** (2006) 347
9. Kahn, C.H.: The verb" be" in ancient greek. (1973)
10. Mimānsak, Y.: Sanskrit Vyakaran Shastra ka Itihas. Yudhishthir Mimānsak (1985)
11. Staal, J.F., Staal, F.: A reader on the Sanskrit grammarians. MIT Press Cambridge, MA (1972)
12. Potter, K.H.: The Encyclopedia of Indian Philosophies, Volume 3: Advaita Vedanta Up to Samkara and His Pupils. Volume 3. Princeton University Press (2014)
13. Bhattacharya, B.: Yāska's Nirukta and the Science of Etymology: An Historical and Critical Survey. Firma KL Mukhopadhyay (1958)
14. Ogawa, H.: Process & language: A study of the mahabha, sya ad a1. 3.1 bhuvadayo dhatavah. Motilal Banarsidass, Delhi, (2005)
15. Leibniz, G.W.: The labyrinth of the continuum: Writings on the continuum problem, 1672-1686. (2001)
16. Brentano, F.: Philosophical Investigations on Time, Space and the Continuum (Routledge Revivals). Routledge (2009)
17. Pantulu, J.: Sri Suryarayandhra nighantuvu. Andhra Pradesh Sahitya Akademi (1936)
18. Bharati, A., Sangal, R.: Computational paninian grammar framework. Supertagging: Using Complex Lexical Descriptions in Natural Language Processing (2007) 355

19. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational linguistics **22** (1996) 249–254
20. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics (1977) 159–174
21. Leeuwenberg, A., Vela, M., Dehdari, J., van Genabith, J.: A minimally supervised approach for synonym extraction with word embeddings. The Prague Bulletin of Mathematical Linguistics **105** (2016) 111–142
22. Singh, S.B.R.P.D., Bhattacharyya, P.: Merging verb senses of hindi wordnet using word embeddings. In: 11th International Conference on Natural Language Processing. (2014) 344
23. : LAC hindi corpus (2014) LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119
25. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, Association for Computational Linguistics (2000) 1–6