A Statistical Study of Telugu Treebanks

PRAVEEN GATLA

Abstract

The paper is an attempt to compare Hyderabad Telugu Treebank (HTTB) and HCU-IIIT-H Telugu Treebank from a statistical point of view. HTTB has 2,715 annotated sentences and HCU-IIIT-H TTB has 3,222 annotated sentences. Both the Treebanks were annotated by following Paninian Grammar Formalism proposed by Bharati, A.; Sharma, D.M.; Husain, S.; Bai, L.; Begam, R. and Sangal, R. (2009). *HTTB is an inter-chunk-based treebank data. HCU-IIIT-H TTB is an intra-chunk-based treebank data. Both the treebanks' data size is random. Later, the paper discusses the Telugu Treebanks in detail. The paper focuses on statistical frequencies viz. POS, Chunk and Syntactic labels. VM (3807 times) and NN (5486 times) are the frequent POS labels in HTTB and HCU-IIIT-H TTB respectively. NP (7954 and 6223 times) is the frequent phrasal category in both the treebanks. The most frequent k-labels are kartā(k1) (2375-2381 times) and karma(k2) (1408-1437 times) and non-frequent label is karaṇa(k3) (17-39 times) in both the treebanks. The most frequent non-k-labels are verb modifier (vmod) (949 times) and noun modifier (nmod) (1033 times) in both the treebanks. The statistical distribution mentions the coverage of the labels (kāraka, non-kāraka) of both the Telugu treebanks. Later it discusses the comparison of both the treebanks and tries to provide the reasons for the highest and lowest frequencies in both the treebanks. k1 and k2 have 60% of the coverage in karaka labels, vmod, nmod, adv, ccof, pof also has 60% of the coverage in non-karaka labels. This kind of statistical study can help to boost the accuracy of the parser.*

Keywords: Treebank, Paninian Grammar, Telugu, *kāraka*, non-*kāraka*, Statistical Frequency, Coverage, Parser.

1. Introduction

The creation of language resources is one of the most challenging tasks in the field of Natural Language Processing. One needs to read and understand the natural language text by making use of one's intuition as a native speaker and his linguistic knowledge. It requires a lot of training in the field of language and linguistics to encode linguistic information. Based on that, treebanks can be created for Indian languages. A plain or simple text, which is encoded with linguistic information, is called annotated data. This kind of lexical resource is useful to develop syntactic parsers for Indian languages. Such tasks involve huge human resources, time, and financial support. In the past, treebanks were created for English and other languages based on different grammatical formalisms (Phrase Structure Grammar, Dependency Grammar, Paninian Grammar, Context-Free Grammar, Universal Dependency Grammar) which are Penn Treebank (Marcus, M.*;* Santorini, B. & Marcinkiewicz, M. A. 1993), Prague Dependency Treebank (Hajičová 1998) so on. Each Grammar Formalism has its own limitations. In order to create treebanks, researchers have used Paninian Grammar, Dependency Grammar, and Universal Grammar Formalisms, which have helped to create syntactic parsers for Indian languages (Hindi, Telugu, Tamil, Marathi, Bangla). The main goal of the present research paper is to compare the two Telugu treebanks. They are HTTB[1] (Praveen 2019) and HCU-IIIT-H TTB[2] (Nallani, S.; Shrivastava, M*;* & Sharma, D. 2020). Praveen (2019) has created 2,715 (sentences) Telugu treebank data by following Paninian Grammar Formalism. Apart from this, HCU-IIIT-H has developed 3,222 Telugu treebank data (sentences) based on Paninian Grammar Formalism. We

[1] Hyderabad Telugu Treebank.
[2] HCU-IIIT-H Telugu Treebank.

considered these two Telugu treebanks for the statistical study. In this paper, we compare both the treebanks from the statistical point of view and try to identify the average number of words per sentence, statistical frequency of Parts of Speech (POS) categories, statistical frequency of Chunks (Phrases), statistical frequency of Telugu treebanks data (*kāraka* labels and non-*kāraka* labels). The paper is organized into five sections. Section 2 discusses related works on treebanks (Indian Languages). Section 3 presents a brief overview of Telugu treebanks, Section 4 describes the statistical frequency of *kāraka and non- kāraka* labels of Telugu treebanks, the significance of the *kāraka* and *non-kāraka* labels in both the Telugu treebanks. Finally, we conclude our paper in Section 5.

2. Related Works

In this section, we discuss some of the relevant research works on treebanks. Treebanks have been developed by following different grammar formalisms. They are Phrase Structure Grammar, Paninian Grammar, Context-Free Grammar, Dependency Grammar, Universal Dependency Grammar. Marcus (1993) describes the construction of a large annotated corpus which is named as Penn Treebank. This resource was developed as a part of the Penn Treebank Project. It was a three-year project from 1989 to 1992. This corpus consists of POS information and skeletal syntactic structure (partially). Penn Treebank is a good resource for linguistic theory (Robert Ingria) and psychological modeling (Niv 1991). Penn Treebank has been extended to other languages like Chinese, Arabic, French, Spanish, etc. Begum, R.*;* Husain, S.*;* Bai, L. and Sharma, D. M. (2008) made an attempt to create Hindi annotated data using the Paninian Grammatical model for the first time. They have annotated almost a million words (nearly 1403 sentences) of Hindi corpus. In this framework, twenty-eight relations were considered for the annotation. It consists

of six basic *kāraka*'s. They are *adhikaraṇa* (k7) 'location', *apaadaan* (k5) 'source', *sampradaan* (k4) 'recipient', *karaṇa* (k3) 'instrument', *karma* (k2) 'theme', *kartā* (k1) 'agent'. Bharati, A.*;* Gupta, M., Yadav, V., Gali, K., and Sharma, D. M. (2009) proposed a simple parser for Indian languages in a dependency framework. They describe a syntactic parser, which follows a grammar-driven approach. They described a grammar-oriented model that makes use of linguistic features to identify relations.

The proposed parser was modeled based on the Paninian grammatical approach. They have shown that with the help of robust rules one can achieve high performance in the identification of various levels of dependency relations. Bhatt, R.*;* Narasimhan, B.*;* Palmer, M.*;* Rambow, O.*;* Sharma, D.M. and Xia, F. (2009) discusses multi-representational and multi-layered treebank. They discuss the multi-representational treebank which provides clues for syntactic dependency version and phrase structure version based on the DS (Dependency Structure) and PS (Phrase-Structure) guidelines. They have developed this treebank based on PropBank and predicate-argument annotation. This approach anticipates that the addition of the PropBank annotation to Dependency Structure (DS) will provide a rich and adequate amount of structure for PS conversion. De, S.*;* Dhar, A. and Garain, U. (2009) have worked on Bangla parsing by following constraint-based dependency parsing. They have used 1000 Bangla annotated sentences to train the system. Chatterji, S.*;* Sonare, P.*;* Sarkar, S and Roy, D. (2009) proposed a hybrid-based approach to parse Bengali sentences. The system tried to work on data-driven dependency parsers. Shailaja (2009) has developed simple Sanskrit sentences rule-based parser. The CLIPS expert system was used to formulate the rules. The developed parser can handle *kāraka* and *upapada vibhakti* relations. Fifteen rules were formulated to handle different

types of Sanskrit sentences. This attempt was the preliminary attempt to develop the Sanskrit parser. Kulakarni (2010) has made a formal attempt to explore the *kāraka* relations in Sanskrit by using Paninian Grammar Formalism. The main attempt is to identify the various *kāraka* relations between the words to extract only syntactic-semantic relations which depend on linguistic or grammatical information in a sentence.

As a part of it, they have annotated 110 (525 tokens) simple sentences which have a single finite verb. The average length of the sentence is 5 words and the maximum length of the sentence is 14 words. Among 110 sentences, 97 sentences output was correct and the remaining 13 sentences were wrongly parsed. Kulakarni and Ramakrishnamacharyulu (2013) discussed some of the specific issues in parsing the Sanskrit texts. In this work, they tried to handle different kinds of constructions in Sanskrit. They are *abhihita*, indeclinables (*avyaya*), inter-sentential connectives, anaphora, conjunctions, and disjunctions. Gade (2014) has worked out on two different treebanks' (Hindi and Sanskrit). She has considered 2300 sentences manually and extracted 1800 sentences which are released for the ICON-2009 Tool Contest (Hindi and Sanskrit) (Husain, S.*;* Mannem, P.*;* Ambati, B.R. and Gadde, P. 2010). Vempaty, C.*;* Naidu, V.*;* Husain, S.*;* Kiran, R.*;* Bai, L.*;* Sharma, D.M., and Sangal, R.*;* (2010) were the first attempt to create Telugu Treebank at LTRC, IIIT-H[3]. They manually annotated 1457 Telugu sentences by following Paninian Grammar Formalism (Bharati, A.*;* Sharma, D.M.*;* Husain, S.*;* Bai, L.*;* Begam, R. and Sangal, R. 2009). Later as a part of IL-IL MT[4] project (Phase II) funded by the Ministry of Electronics and Information Technology (MeitY)[5],

[3] Language Technology Research Center, International Institute of Information Technology, Hyderabad.

[4] Indian Languages to Indian Languages Machine Translation Systems

[5] https://meity.gov.in/

Government of India, it was decided to develop a simple syntactic parser for the nine Indian languages. As a part of this task, a group led by Umamaheshwar Rao (2010-2016) have developed 5,000 (sentences) HCU Telugu treebank at CALTS, UoH[6].

Recently, by combining IIIT-H Telugu treebank consisting of 1600 sentences from ICON 2009 tools contest, 200 (sentences) Telugu treebank data (IIIT-H), and 5,000 (sentences) HCU Telugu Treebank (Umamaheshwar Rao, G.*;* Koppaka, R.*;* Addanki, S. 2012; Rajyarama & Srinivas, 2015a & 2015b) have been combined into one set. Nallani, S.; Shrivastava, M, and Sharma, D.M. (2020) have formatted, cleaned and released the licensed final Telugu treebank data consisting of 3,222 sentences under the Creative Commons License Attribution Noncomercial Share 4.0.1. International. International. Rama and Soumya (2017) have worked on Telugu treebank. They have followed the Universal Dependency framework and annotated 1328 sentences from Telugu grammar. The treebank developed by them is freely available at Universal Dependencies[7] version 2.1. They discussed corpus annotation, parts-of-speech annotation, morphology, Universal Dependency relations in their paper. They also reported the preliminary tagging and parsing results with UDPipe. Apart from that, Universal Dependency treebanks have been developed for nine Indian languages. They are Bhojpuri, Hindi, Hindi-English, Kangri, Magahi, Marathi, Sanskrit, Tamil, and Urdu. There are four upcoming Indian languages under UD treebank. They are Bengali, Assamese, Kannada, Pnar[8].

[6] Centre for Applied Linguistics and Translation Studies, University of Hyderabad.

[7] https://github.com/UniversalDependencies/UD_Telugu-MTG

8https://universaldependencies.org/

3. Telugu Treebanks

Telugu is a Dravidian language. It is a morphologically rich language. A series of suffixes can be attached to a single root in Telugu. In the development of the treebanks, we give much more importance to the morphological (inflection) information because it gives gender, number, person, case information for nouns, tense, aspect, modality information for verbs. All these interpretations reflect at the morphological level in Telugu. Here, we consider two Telugu treebanks which are developed based on Paninian Grammar Formalism for statistical study. Based on this formalism, dependency structure (DS) guidelines were developed by Akshar Bharati group[9]. As a part of IL-IL MT project[10], this group developed the annotation guidelines to create treebanks for Indian languages. The baseline for creating these guidelines is Paninian grammar. In this framework[11], a sentence is considered as one unit where the verb is the central notion. Apart from that, other constituents also play an important role in a sentence. The *kāraka relations* denote syntactico-semantic relations between the verb and other constituents in a sentence (Cf. Sangal, R.*;* Chaitanya, V. & Bharati, A. 1995). There are two types of relations in this scheme i.e., *kāraka* and non-*kāraka*. The *kāraka* relations are *kartā* (k1) 'doer', *karma* (k2) 'object', *karaṇa* (k3) 'instrument', *saṃpradāna* (k4) 'receiver', *apādāna* (k5) 'source', *adhikaraṇa* (k7) 'location' and non-*kāraka* relations are *ṣaṣṭhī (*r6) (genitive, possessive), *hētuḥ* (rh) ‘reason’, *tādarthya* (rt) ‘purpose’, *adjectival* modifiers (jjmod) and *adverbial modifiers* (rbmod) etc. Based on the types of relations, the

[9] AnnCorra: Tree Banks for Indian Languages Guidelines for Annotating Hindi Treebank (Ver 2.0).

[10] Indian Languages to Indian Languages Machine Translation System Project (Phase I and II) funded by Ministry of Electronics and Information Technology, Government of India.

[11] AnnCorra: TreeBanks for Indian Languages.

dependency tags are also classified into two types. They are inter-chunk[12], intra-chunk[13]. The chunks (quasi phrases) are considered as the heads in inter-chunk relations (annotation), whereas in intra-chunk annotation each word or token is marked with a relation. Bharati, A., Sharma, D.M., Husain, S., Bai, L., Begam, R., and Sangal, R. (2009) have developed DS guidelines to create treebanks. In this framework, *kāraka* and non-*kāraka* relations are denoted in a sentence. *kāraka* relations have tags that start with a '*k*' and are followed by a numerical digit (e.g., 1 to 5 and 7). They are *kartā* (k1), *karma* (k2), *karaṇa* (k3), *sampradāna* (k4), *apādāna* (k5), *adhikaraṇa* (k7), etc. These *kāraka*'s are further fine-grained as sub-tags of *kartā* such as *kartā samānādhikaraṇa* (k1s), *prayojya kartā* (Causee; jk1), *prayojaka kartā* (Causer; pk1). The non-*kāraka* tags either begin with '*r*' or '*c*' or '*p*'. They are *ṣaṣṭhī* 'genitive or possessive' (r6), *hētuḥ* (reason) (rh), *tādarthya* 'purpose' (rt), Coordination (ccof), Part of (pof), etc. The different types of dependency relations are mentioned in Bharati, A.; Sharma, D.M.; Husain, S.; Bai, L.; Begam, R. and Sangal, R. (2009) which shows "the relations from coarser level to finer level on a modifier-modified paradigm" (Bharati, A.; Sharma, D.M.; Husain, S.; Bai, L.; Begam, R. and Sangal, R. 2009).

3.1 Hyderabad Telugu Treebank (HTTB)

We have adopted DS Guidelines which are developed by Bharati, A.; Sharma, D.M.; Husain, S.; Bai, L.; Begam, R. and Sangal, R. (2009) which are followed to create HTTB. As a part of it, we have developed 2,715 Telugu treebank data by following DS guidelines in 2009. Sentences are extracted from various sources such as literary texts and grammar books viz. Krishnamurti and Sarma (1968), Krishnamurti and Gwynn

[12] Inter-chunk mark the kāraka relations that occurs between any two chunks.

[13] tags mark the kāraka relations within a chunk.

(1985), Krishnamurti (1991, 2003, 2009), Ramarao (2002, 1975), Subrahmanyam (1984), Ramakrishna Reddy (1986), Subbarao (2012), Usha Rani (1980), Prakasam (2018) to develop the Telugu Treebank data. In this treebank, we have considered Bharati, A.; Sangal, R.; Sharma, D.M. and Bai, L. (2006) POS categories. This tagset consists of 26 POS categories. It is an inter-chunk-based treebank data.

3.2 HCU-IIIT-H Telugu Treebank (HCU-IIIT-H TB)

Nallani, S.; Shrivastava, M and Sharma, D. (2020) combined the HCU TTB and IIIT-H TTB data into one set and made it available for public access. In this, there are 3,222 annotated sentences. It is intra-chunk-based data. Telugu POS tagged data have been converted by following the latest BIS tagset[14] (Bureau for Indian Standards). The BIS tagset is a standardized POS tagging guideline for all Indian languages. This tagset consists of 11 POS categories and most of the categories have further divided into fine-grained POS tags.

4. Statistical Frequency of Telugu Treebanks

In this section, we discuss the statistical frequencies of HTTB and HCU-IIIT-H TTB. The average length of the sentences of HTTB (Cf. Praveen, 2019) and HCU-IIIT-H TTB (Cf. Nallani, S.; Shrivastava, M, and Sharma, D. 2020) are 6 and 5.5 respectively. As a part of this exercise, we have listed out the frequencies of POS categories, phrases (chunks), *kāraka,* and non-*kāraka* labels which are discussed in this section.

4.1 Statistical Frequency of Parts of Speech (POS) Categories of Telugu Treebanks

We have identified the occurrences of each POS category in the Telugu treebanks dataset. Here, we have considered only those POS categories which were used in Bharati, A.*;* Sangal,

[14] http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf

R.; Sharma, D.M. and Bai, L. (2006) for calculating statistical frequency. In HTTB, VM (Main Verb) has occurred 3,807 times (highest) and QO(Ordinal) has occurred 11 times (lowest). In HCU-IIIT-H TTB, NN (Common Noun) has occurred 5486 times (highest) and PSP (Post-position) has occurred 2 times (lowest). In the latest BIS tagset, Nallani, S.; Shrivastava, M and Sharma, D. (2020) have used RD_PUNC for *SYM*, *PR_PRQ* for *WQ* so and so forth. *RDP* was not found in HCU-IIIT-H TTB. Parts of Speech (POS) categories and their frequencies are shown in Table 1. Hyphen denotes not found in Table1.

Sl. No.	**POS Tags**	**HTTB Frequency Count**	**HCU-IIIT-H TTB Frequency Count**
1	VM	3807	4317
2	NN	3509	5486
3	SYM/RD_PUNC	2937	3330
4	PRP	1741	1246
5	NNP	1020	695
6	NST	426	316
7	RB	292	432
8	DEM	261	237
9	WQ/PR_PRQ	246	215
10	JJ	230	414
11	VAUX	175	59
12	CC/CC_CCS/CC_CCD	164	262
13	QC	125	274
14	PSP	101	2
15	RP/RPD	98	65
16	QF	81	191
17	UT/CC_CCS_UT	81	105
18	INTF	46	85
19	RDP	37	-
20	NULL	19	19
21	CL	13	14
22	QO	11	21

Table 1: POS Frequency in the Telugu treebanks

4.2 Statistical Frequency of Chunks (Phrases) of Telugu Treebanks

We have identified the occurrences of each Phrasal (Chunks) category in Telugu treebanks dataset. We counted the frequency of each Phrasal category and their frequency in Table 2. Because HCU-IIIT-H TTB is available in the intra-chunk format. We considered only chunk(phrasal) heads for the frequency count. Hyphen denotes not found in Table2.

Sl. No.	Phrasal Categories	HTTB Frequency Count	HCU-IIIT TTB Frequency Count
1	NP	6223	7954
2	VGF	3739	3314
3	VGNF	997	865
4	RBP	170	458
5	VGNN	124	126
6	BLK	103	200
7	CCP	162	317
8	JJP	7	103
9	VGINF	-	6

Table 2: Chunk frequency in the Telugu treebanks

Noun Phrase has occurred 7954 and 6223 times in the HTTB and HCU-IIIT-H TTB respectively. It is the highest frequent phrasal category in both the treebanks. The lowest frequent phrasal category is JJ (Adjectival Phrase), VGINF (Infinitival Verbal Phrase) have occurred 7 and 6 times in HTTB and HCU-IIIT-H TTB respectively.

4.3 Statistical Frequency of Telugu Treebanks (Labels)

We have considered both the Telugu treebanks[15]. By using this annotated data, we have calculated the *k-labels* and *non-k-labels* statistical frequency separately and discussed them in detail.

[15] HTTB and HCU-IIIT-H TTB

4.3.1 Statistical Frequency of Hyderabad Telugu Treebank

In HTTB, the highest and lowest frequency for *kāraka* labels are *kartā*(k1) *kāraka* 2375, *karaṇa*(k3) *kāraka* 17 times respectively. Similarly, the highest and lowest frequency for *non-kāraka* labels are Verb Modifier (vmod) 949 times and *associative (ras-k2)* 4 times respectively. Figure 1 and 2 represents the statistical frequency of HTTB viz. *kāraka* labels and non-*kāraka* labels separately.

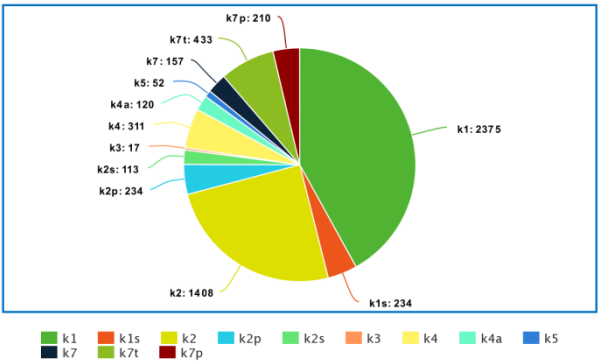

Figure 1: Statistical frequency of *kāraka* labels of HTTB

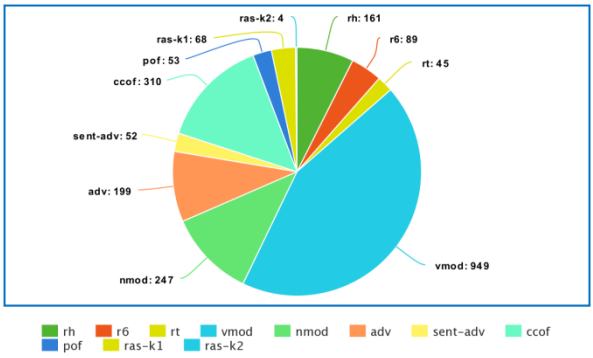

Figure 2: Statistical frequency of *non-kāraka* labels of HTTB

4.3.2 Statistical Frequency of HCU-IIIT-H Telugu Treebank

In HCU-IIIT-H TTB, the highest and lowest frequency for *kāraka* labels are *kartā*(k1) 2381, *karma samānādhikarana* (k2s) 32 times respectively. Similarly, the highest and lowest frequency for *non-kāraka* labels is *Noun Modifier* (nmod) 1033, ras-k2 (Relation for Associative) 6 times. Figure 3 and 4 represents the statistical frequency of HCU-IIIT-H TTB viz. *kāraka* and *non-kāraka* labels separately.

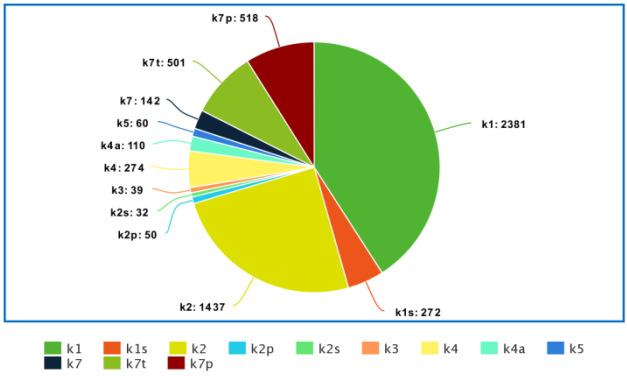

Figure 3: Statistical frequency of *kāraka* labels of HCU-IIIT-H TTB

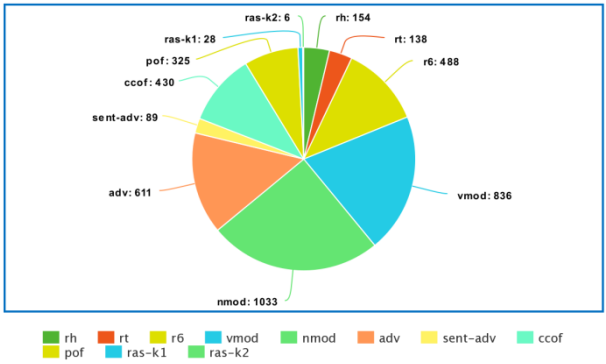

Figure 4: Statistical frequency of *non-kāraka* labels of HCU-IIIT-H TTB

4.4 Statistical Comparison of HTTB and HCU-IIIT-H TTB

Here, we try to compare both the Telugu treebanks statistics one by one based on the statistical frequency. The comparison of both the treebanks can be seen in Figure 5. We try to draw our observations based *on the highest* and lowest frequency of kāraka and non-kāraka labels. Here, one more important point is that the size of the Telugu treebanks is not the same. They are random in size. Statistical frequency of both the treebanks (HTTB and HCU-IIIT-H TTB) viz. k1 label is 2375-2381, k1s is 234-272, k2 is 1408-1437, k2p is 234-50, k2s is 113-32, k3 is 17-39, k4 is 311-274, k4a is 120-110, k5 is 52-60, k7 is 157-142, k7t is 433-501, k7p is 210-518, rh is 161-154, r6 is 89-488, rt is 45-138, vmod is 949-836, nmod is 247-1033, adv is 199-611, sent-adv is 52-89, pof is 53-325, ccof is 310-430, ras-k1 is 68-28, ras-k2 is 4-6 etc[16] (See Figure 1, 2, 3 and 4). Based on the two Telugu treebanks, one can predict that most of the time a sentence would have kartā(k1), *karma*(k2) for sure. It means *that the subject* and object are mandatory in a sentence. Another observation is that *saṃpradāna*(k4), apādāna(k5), *adhikaraṇa*(k7) are expected in a sentence optionally. I*n comparison with kartā(k1) and karma(k2), their (k4, k5, k7)* occurrences or frequencies are very less in both the treebanks. When we compare *non-kāraka* relations or labels viz. hētu*(rh), verb modifiers (vmod), noun modifiers (nmod), coordination (ccof),* sent-adv *(Sentential Adverbs)* frequencies are higher than *tādarthya*(rt), *ṣaṣṭhī*(r6), adverbs*(adv), pof* (Part of), *ras-k1 (Associative with* kartā), ras-k2 (Associative with karma*)*. There are certain labels, which *have a* drastic difference in the frequency count. For example, *r6* is 89-488, *rt* is 45-138, *nmod* 247-1033, adv 199-611, pof 53-325, ras-k1 68-28. HTTB data has a low frequency (treebank labels) in comparison with HCU-IIIT-HTTB.

[16] These numbers denote number of occurrences in HTTB and HCU-IIIT-H Telugu treebanks.

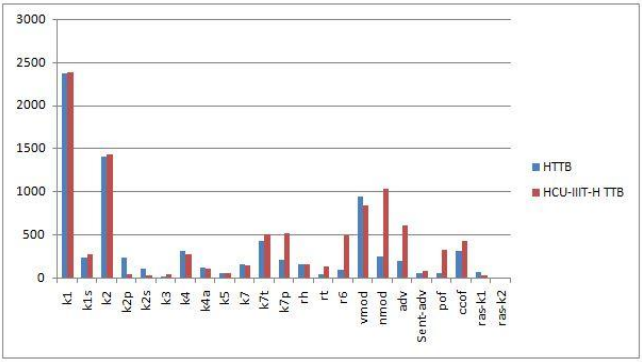

Figure 5: Comparison of HTTB and HCU-IIIT-H TTB labels (*kāraka* and *non-kāraka* labels)

4.5 Significance of the Distribution

The statistical distribution of both the Telugu treebanks is discussed in detail with their coverage in this section. Here HTTB and HCU-IIIT-H TTB coverage of *kāraka* and *non-kāraka* labels are given one by one. k1 is 41.9%-40.9%, k1s is 4.1%-4.7%, k2 is 24.9%-24.7%, k2p is 4.1%-0.9%, k2s is 2.0%-0.6%, k3 is 0.3%-0.7%, k4 is 5.5%-4.7%, k4a is 2.1%-1.9%, k5 is 0.9%-1.0%, k7 is 2.8%-2.4%, k7t is 7.6%-8.6%, k7p is 3.7%-8.9%, rh is 7.4%-3.7%, r6 is 4.1%-11.8%, rt is 2.1%-3.3%, vmod is 43.6%-20.2%, nmod is 11.3%-25.0%, adv is 9.1%-14.8%, sent-adv is 2.4%-2.2%, ccof is 14.2%-10.4%, pof is 2.4%-7.9%, ras-k1 is 3.1%-0.7%, ras-k2 is 0.2%-0.1% in HTTB and HCU-IIIT-H TTB respectively. Here we have considered *kāraka* and *non-kāraka* labels separately for calculating the coverage.

In Telugu, there is no overt distinction between *kartā* (k1) and *karma* (k2) at the syntactic level. Because most of the time k1 and k2 are marked with zero markers (null). Sometimes k1 is also realized with '*ki*', '*ceta*' vibhakti markers. In the same way, k2 is expressed with the '*ni*/*0*' vibhakti marker. The vibhakti

marker '*ni*' is mandatory for animate nouns and it is optional for inanimate nouns in Telugu. Apart from that semantic information is mandatory to recognize or parse k1 and k2 correctly. In other words, the animate and inanimate distinction should be made in the Treebank data to recognize k1 and k2 correctly. Otherwise, there is a chance of recognizing k1 as k2 and k2 as k1. It may lead to incorrect parsing. By comparing the coverages of the Telugu treebanks, we can say that 65% of the TTB data is covered by k1 and k2 approximately and the remaining 35% of the TTB data is covered by remaining *kāraka* labels such as k1s, k2p, k2s, k3, k4, k4a, k5, k7, k7t, k7p, etc. In the case of *non-kāraka* labels, vmod, nmod, adv, and ccof covers more than 60% of the data, and the remaining 40% covered by rh, r6, rt, sent-adv, pof, ras-k1, ras-k2 etc. It is observed that k1, k2, vmod, nmod, adv, ccof labels are more important during the creation of the annotated data (Treebank data). The coverage of these four labels is more than 60%. Among k-labels, k3 has 0.3%-0.7% of the coverage in both the treebanks. The probable reason could be the case marker '-to' ('with') which denotes an instrumental case. But the same case marker '-to' also denotes ras-k1 which means relation for associative with *kartā*. For example, *ravi kṛṣṇa* ***tō*** *bajāruku veḷlāḍu* 'Ravi went to market with Krishna'. Here, 'with Krishna' is marked as ras-k1 but not as k3. Because it does not denote the instrumental case (*karaṇa kāraka*). Another reason could be the lack of such constructions in both the Telugu treebanks database. k4 has 5.5%-4.7% of the coverage in both the Telugu treebanks. The relation k4 is one of the most ambiguous *kāraka* relations. The dative case marker '-ki' is used to denote the k4 (*saṃpradāna*) relation generally. But it also denotes the *hētu* (rh) 'reason' relation (*non-kāraka* relation). For example, *āme doṃga****ki*** *bhayapaḍiṃdi* 'She feared because of thief'. Here 'because of' is interpreted as *hētu* (rh) 'reason' but not as *saṃpradāna* (k4) 'receiver'. The case

marker '*-ki*' is the most ambiguous in Telugu. Because of it, the maximum coverage of k4 is 5.5% and a minimum 4.7% respectively in both TTB's. k5 (source) has 0.9%-1.0% coverage only in both the treebanks. The case marker '*nuṃḍi*' is used to denote the k5 relation. Syntactic constructions which denote k5 relation might be less in the database. It is an unambiguous relation in Telugu. Hierarchically k7 '*viśayādhikaraṇa*' is the main *kāraka* relation and k7t and k7p are the sub-tags of k7. All these three relations are denoted with nouns with space and time (NST). Among these three (k7, k7t, k7p), k7p has 3.7%-8.9% coverage, k7t has 7.6%-8.6% coverage, k7 has 2.8%-2.4% in both the TTB's. NST's (POS tags) have occurred 426, 316 times in both the treebanks. Among all the phrasal categories, NP (Noun Phrase) is the highest frequent phrasal category which has occurred 6223, 7954 times respectively (See Section 4.1 and 4.2). By looking at these frequencies, it is quite natural that some NST's are expected in the natural language. Generally, they are spatial and temporal nouns. In HTTB, k7t has the highest coverage (7.6%). In HCU-IIIT-H TTB k7p has the highest coverage (8.9%).

Among *non-kāraka* relations, vmod, nmod, coordination (ccof), *sent-adv* (Sentential Adverbs) have the highest coverage. Naturally, Paninian Grammar Formalism expresses modifier-modified relations. Among *non-kāraka* relations, vmod's, nmod's, adverbs, genitives, part of relations (pof) (complex predicates) are large in number. It means that there might be a large number of modifiers that precede the nouns and verbs respectively in the annotated data (both the TTB). They are vmod 43.6%-20.2%, nmod 11.7%-25.0%, adv 9.1%-14.8%, r6 4.1%-11.8%, pof 2.4%-7.9% in which most of them are having the highest coverage. The remaining *non-kāraka* relations rh, rt, sent-adv, ras-k1, ras-k2 are having below 8.0% of the coverage in both TTB's.

5. Conclusion

The present paper is an attempt to compare two Telugu treebanks (HTTB and HCU-IIIT-H TTB). HTTB consists of 2,715 and HCU-IIIT-H TTB consists of 3,222 annotated sentences. Both the treebanks have been created by following DS guidelines which are developed by Bharati, A.*;* Sharma, D. M.*;* Husain, S.*;* Bai, L.*;* Begam, R. and Sangal, R. (2009). The statistical study has been done at three levels. They are POS, chunk (Phrase), and dependency labels (*kāraka* and *non-kāraka*). HTTB sentence length is lesser than HCU-IIIT-H TTB. HTTB has less number of nouns than HCU-IIIT-H TTB at the POS level. Telugu sentences are extracted from Telugu grammars to build HTTB data whereas HCU-IIIT-H TTB data has been extracted from tourism and health domain as a part of IL-IL MT project. In the present study, we found that VM (3,807 times) and NN (5486 times) are the highest frequent POS categories and QO (11 times), PSP (2 times) are lowest frequent POS categories in HTTB and HCU-IIIT-H TTB respectively. Similarly, NP (7954 and 6223 times) is the highest frequent phrasal category in both the treebanks whereas JJP (7 times), VGINF (6 times) are the lowest frequent phrasal categories in both the treebanks respectively. The major observation is that 65% of the Telugu treebank data is covered by k1 and k2 (*kāraka* relations). The remaining 35% is covered by k1s, k2p, k2s, k3, k4, k4a, k5, k7, k7t, k7p. In *non-kāraka* relations, vmod, nmod, adv and ccof has more than 60% of the coverage. The remaining 40% is covered by rh, rt, sent-adv, pof, ras-k1, ras-k2 etc. Based on these two statistics, the major coverage is for k1, k2, vmod, nmod, adv, ccof, pof. The coverage of these four labels is more than 60%. This kind of findings will help to make the generalizations based on the statistical frequencies of the treebanks. These generalizations will help the annotator to concentrate on highest frequent labels instead of the lowest frequent labels during the treebank

validation. In both the TTB's, *k1, k2, vmod, nmod, ccof, sent-adv, genitives, pof* can be crosschecked or validated for accurate Treebank data. By doing such a kind of statistical study one can know where to spend or concentrate or devout time to improve the treebank data (annotated data). This kind of statistical study is useful to train the human annotators to create the Treebank data. It also helps to boost the accuracy rate of the parsers.

Acknowledgment

I thank Prof. G. Umamaheshwar Rao who was motivated to create Hyderabad Telugu Treebank. I thank Mr. Y. Vishwanath Naidu and Dr. V. Subrahmanya Madhav Sharma for their valuable discussions to improve the Hyderabad Telugu treebank annotation. I would also like to thank Sneha Nallani, Manish Shrivastava, Dipti Misra Sharma for making the availability of Telugu Treebank (intra-chunk) data for public access under CCLA. Because of it, we could do the statistical study of both the treebanks. I thank the reviewers for their critical comments and suggestions, which helped us to improve the paper.

References

SANGAL, R., CHAITANYA V. & A. BHARATI. 1995. *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice-Hall.

BHARATI, A., R SANGAL, D. M. SHARMA & L. BAI. 2006. *Anncorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages*. Hyderabad: IIIT-TR31. 1-38.

BHARATI, A., D. M. SHARMA, S. HUSAIN, L. BAI, R. BEGAM, & R. SANGAL. 2009. *AnnCorra: TreeBanks for Indian Languages, Guidelines for Annotating Hindi TreeBank* (version-2.0). Hyderabad: IIIT.

BHARATI, A., M. GUPTA, V. YADAV, K. GALI, & D. M. SHARMA. 2009. Simple Parser for Indian Languages in a Dependency Framework. *Proceedings of the Third Linguistic Annotation Workshop*. 162–165. US: Association for Computational Linguistics.

BHATT, R., B. NARASIMHAN, M. PALMER, O. RAMBOW, D. M. SHARMA & F. XIA. 2009. A Multi-representational and Multi-layered Treebank for Hindi/Urdu. *Proceedings of the Third Linguistic Annotation Workshop (LAW III).* 186-189. Singapore: World Scientific Publishing Co Pvt Ltd.

CHATTERJI, S., P. SONARE, S. SARKAR & D. ROY. 2009. Grammar Driven Rules for Hybrid Bengali Dependency Parsing. *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing*. Hyderabad: IIIT.

DE, S., A. DHAR & U. GARAIN. 2009. Structure Simplification and Demand Satisfaction Approach to Dependency Parsing for Bangla. *Proceedings of 6th International Conference on Natural Language Processing (ICON) Tool Contest: Indian Language Dependency Parsing*. 25-31. Hyderabad: IIIT.

GADE, R. P. 2014. Dependency Parsing Approaches for Indian Languages: Hindi and Sanskrit. Hyderabad: IIIT doctoral dissertation.

HAJIČOVÁ, E. 1998. *Prague Dependency Treebank: From Analytic to Tectogrammatical Annotations*. *Proceedings of 2nd TST.* 45-50. Brno, New York: Springer-Verlag Berlin Heidelberg.

KRISHNAMURTI, BH. 1991. *Studies in Dravidian and General Linguistics: A Festschrift for Bh. Krishnamurti* 6. Hyderabad: Osmania University.

KRISHNAMURTI, BH. 2003. *The Dravidian Languages*. Cambridge: Cambridge University Press.

KRISHNAMURTI, BH. 2009. *Studies in Telugu Linguistics*. Hyderabad: C.P. Brown Academy.

KRISHNAMURTI, BH. & J. P. L. GWYN. 1985. *A Grammar of Modern Telugu*. Delhi: OUP.

KRISHNAMURTI, BH. & P. S. SARMA. 1968. *A Basic Course in Modern Telugu*. Delhi: Motilal Banarsidass.

KULKARNI, A., S. POKAR & D. SHUKL. 2010. Designing a Constraint Based Parser for Sanskrit. In *International Sanskrit Computational Linguistics Symposium*. 70-90. Berlin: Springer.

KULKARNI, A. & K. RAMAKRISHNAMACHARYULU. 2013. *Parsing Sanskrit Texts: Some Relation Specific Issues*. In Chaitali Dangarikar & Malhar Kulkarni (ed.), *Proceedings of the 5th International Sanskrit Computational Linguistics Symposium*. New Delhi: DK Printworld (P) Ltd.

MARCUS, M., B. SANTORINI & M. A. MARCINKIEWICZ. 1993. *Building a Large Annotated Corpus of English: The Penn Treebank.* University of Pennsylvania. Technical Report No. MS-CIS-93-87. *Computational Linguistics* 19(2). 313-330.

NALLANI, S., M. SHRIVASTAVA & D. SHARMA. 2020. *A Fully Expanded Dependency Treebank for Telugu*. *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation.* Marseille: Language Resources and Evaluation Conference (LREC 2020). 39-44.

NIVRE, J., M. C. DE MARNEFFE, F. GINTER, Y. GOLDBERG, J. HAJIC, C. D. MANNING, R. MCDONALD, S. PETROV, S. PYYSALO, N. SILVEIRA & R. TSARFATY. 2016. *Universal Dependencies v1: A Multilingual Treebank Collection*. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1659-1666.

PRAKASAM, V. 2018. *Aadhunika Telugu Vennelakanti Vyaakaranam*. Hyderabad: Nirrvitha Publishing.

PRAVEEN, G. 2019. *Developing a Parser for Telugu*. Hyderabad: HCU doctoral dissertation.

RAJYARAMA, K. & A. SRINIVAS. 2015a. Issues in Developing a Dependency Parser for Telugu: A Linguistic Account. *BHASHAA: An International Journal of Telugu Linguistics*. 93-101. Magazine-4. Hyderabad: Telugu Linguists' Forum.

RAJYARAMA, K. & A. SRINIVAS. 2015b. Yantranuvadamlo Ani. *BHASHAA: An International Journal of Telugu Linguistics.* 35-45. Magazine-1. Hyderabad: Telugu Linguists' Forum.

RAMA, T. & S. VAJJALA. 2018. A Dependency Treebank for Telugu. *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16).* 119-128. Prague: Czech Republic.

RAMARAO, C. 1975. *Telugu Vakyam*. Hyderabad: Andhra Pradesh Sahitya Academy.

RAMARAO, C. 2002. *Quest of Subject in Telugu*. In Swarajya Lakshmi (ed.), Case for Language Studies 30. 153-156. Hyderabad: Book Links Corporation.

SHAILAJA, N. 2009. Parser *for Simple Sanskrit Sentences.* Hyderabad: HCU M.Phil dissertation.

SRINIVAS, A. 2012. *Telugu Bhasha-Vyakaranam*. Hyderabad: Akruthi Offset Printers.

SRINIVAS, A. & K. RAJYARAMA. 2015. Hindi-Telugu Yantranuvadamlo Yokka (sheṣa ṣaṣṭhi): Alopa Sutralu. *BHAASHA: International Journal of Telugu Linguistics* 4. 58-67. Hyderabad: Telugu Linguists' Forum.

SUBBARAO, K. V. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge: CUP.

SUBRAHMANYAM, P. S. 1984. *Adhunika Bashashastra Sidhantalu*. Hyderabad: Andhra Pradesh Sahitya Academy.

TESNIÈRE, L. 1959. Eléments de Syntaxe Structurale. (Timothy Osborne & Sylvain Kahane, Trans.). Amsterdam: John Benjamins.

UMAMAHESHWAR RAO, G., R. KOPPAKA, & S. ADDANKI. 2012. Dative Case in Telugu: A Parsing Perspective. *Proceedings*

of the Workshop on Machine Translation and Parsing in Indian Languages. 123-132. Mumbai: COLING.

USHA RANI, A. 1980. *Relativization in Telugu*. Hyderabad: Osmania University doctoral dissertation.

VEMPATY, C., V. NAIDU, S. HUSAIN, R. KIRAN, L. BAI, D. M. SHARMA & R. SANGAL. 2010. Issues in Analyzing Telugu Sentences Towards Building a Telugu Treebank. *International Conference on Intelligent Text Processing and Computational Linguistics*. 50-59. New York: Springer.

***

Cite This Work:

GATLA, PRAVEEN. 2021. A Statistical Study of Telugu Treebanks. *Translation Today*, Vol. 15(1). 145-167.
DOI:10.46623/tt/2021.15.1.ar6