A Dependency Treebank for Telugu

Taraka Rama
University of Oslo

Sowmya Vajjala
Iowa State University, sowmya@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/engl_conf

Recommended Citation
https://lib.dr.iastate.edu/engl_conf/8

This Conference Proceeding is brought to you for free and open access by the English at Iowa State University Digital Repository. It has been accepted for inclusion in English Conference Papers, Posters and Proceedings by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
A Dependency Treebank for Telugu

Abstract
In this paper, we describe the annotation and development of Telugu treebank following the Universal Dependencies framework. We manually annotated 1328 sentences from a Telugu grammar textbook and the treebank is freely available from Universal Dependencies version 2.1.1 In this paper, we discuss some language specific annotation issues and decisions; and report preliminary experiments with POS tagging and dependency parsing. To the best of our knowledge, this is the first freely accessible and open dependency treebank for Telugu.

Keywords
Telugu, Universal dependencies, adverbial clauses, relative clauses

Disciplines
Communication Technology and New Media

Comments
The following proceeding is published as Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 119–128, Prague, Czech Republic, January 23–24, 2018. Distributed under a CC-BY 4.0 licence.

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.
A Dependency Treebank for Telugu

Taraka Rama
Department of Informatics
University of Oslo, Norway
tarakark@ifi.uio.no

Sowmya Vajjala
Applied Linguistics and Technology Program
Iowa State University, USA
sowmya@iastate.edu

Abstract

In this paper, we describe the annotation and development of Telugu treebank following the Universal Dependencies framework. We manually annotated 1328 sentences from a Telugu grammar textbook and the treebank is freely available from Universal Dependencies version 2.1. In this paper, we discuss some language specific annotation issues and decisions; and report preliminary experiments with POS tagging and dependency parsing. To the best of our knowledge, this is the first freely accessible and open dependency treebank for Telugu.

1 Introduction

An annotated treebank is a pre-requisite for developing computational tools that support deeper language processing for any language. Treebanks are typically created with texts collected from specific genre such as news, fiction, Wikipedia, blogs, and Bible. There also exist treebanks for non-canonical text such as learner data (Berzak et al., 2016; Lee et al., 2017). While these treebanks have been used for the development of natural language parsers and other tools, they may not cover infrequent grammatical structures that do not occur in the specific domain of the training data. Grammar books provide an excellent set of examples for annotated sentences that cover a wide range of syntactic structures as these sentences are chosen to illustrate the interesting and unique features in a language. Additionally, such grammar book based treebanks can also be used to test the coverage of statistical parsers trained with a large amount of data from a specific domain or genre. Further, a typical grammar book features short sentences and allows rapid development of a treebank. Hence, they serve as a good starting point for developing a broad coverage treebank (Cöltekin, 2015).

Telugu is a Dravidian Language native to India with 74 million native speakers with a long history of written and oral literature. Despite some published research on development of part-of-speech taggers (PVS and Karthik, 2007) and a treebank (Vempaty et al., 2010), neither of the resources are publicly available. In this paper, we describe our efforts in developing a publicly available treebank for Telugu that covers a range of syntactic constructions and morphological phenomena. We manually annotated 1328 sentences from the Telugu grammar book by Krishnamurti and Gwynn (1985) with (universal) part-of-speech tags and dependency relations. We followed the Universal Dependencies (Nivre et al., 2016) framework for annotation, as it supports the development of treebanks for new languages through extensive documentation. We also report preliminary POS tagging and dependency parsing results using the treebank data and UDPipe (Straka et al., 2016).

The rest of the paper is organized as follows. We describe the related work in section 2. We provide a brief description of linguistic properties of Telugu in section 3. Then, we describe the corpus and annotation environment in section 4. We describe the annotation decisions during the annotation of POS tags in section 5. Section 6 briefly introduces Telugu morphology and is followed by section 7 that discusses dependency relations that are specific to Telugu. We present the results of our POS tagging and parsing experiments in section 8. Finally, we conclude the paper and discuss some directions for future work.

1http://universaldependencies.org/treebanks/te/
2 Related Work

Treebanks for some South Asian languages were developed following the Paninian framework for dependency annotation (Begum et al., 2008) and some of them are publicly available. There were some early efforts towards Telugu dependency treebank development following the Paninian framework (Vempaty et al., 2010). A Telugu treebank was also a part of an Indian language dependency tools contest (Husain et al., 2010). However, none of these resources are publicly available to the best of our knowledge. There have been efforts to convert some of the Indian language treebanks into Universal Dependencies (UD) framework (Tandon et al., 2016) and there is a reasonably large UD treebank for Hindi (Palmer et al., 2009; Bhat et al., 2017). However, except for Tamil (Ramasamy and Žabokrtský, 2012), converted from Prague dependency style (Hajič et al., 2017) to UD, there is no UD treebank for any other language from the Dravidian language family. In this scenario, the availability of free and open UD Telugu treebank would be a good starting point for the future of computational infrastructure support for Telugu and the Dravidian language family.

UD treebanks were developed from scratch for several low-resource languages such as Kazakh (Makazhanov et al.) and Buryat (Badmaeva and Tyers, 2017) in the recent years. Several existing treebanks are also being converted into UD. While there is no Dravidian language in UD other than Tamil, there exist treebanks for other agglutinative languages such as Finnish (Pyysalo et al., 2015), Hungarian (Vincze et al., 2017), Turkish (Çöltekin, 2015; Sulubacak et al., 2016), and Estonian (Muischnek et al., 2016) which provided us with useful insights in dealing with language-specific morphological and syntactic phenomena for Telugu.

3 Telugu

Telugu is one of the 22 languages with official status in India. Telugu belongs to the South-Central subgroup of the Dravidian language family and is mainly spoken in Southern India. The Dravidian language family was the subject of both historical and comparative linguistic research in the later half of twentieth century (Krishnamurti, 2003). Telugu is an agglutinative language like other Dravidian languages such as Tamil or Malayalam. The dominant word order in Telugu is Subject-Object-Verb (SOV) with inclination towards pro-drop. Telugu verbs inflect for gender, number and person. The “be” (vun-, “existential”) verb in Telugu shows agreement with the subject for gender, number, and person. The existential verb has a negative counterpart “not to be” (le-) which can participate in both light and serial verb constructions and also act as the main verb. Telugu does not have a dominant overt coordination as in English or Hindi. Telugu forms subordinate clauses through verbal nouns, verbal adjective, and converbs. Control constructions marked by xcomp relation are rare (less than ten instances in our treebank) in Telugu.

4 Corpus and Annotation

The Telugu treebank currently consists of 1328 sentences and 6465 tokens. The sentences were manually typed in Telugu script (derived from Brahmi script) from the examples in chapters 7–29 in (Krishnamurti and Gwynn, 1985). Many sentences in this book are collected from contemporary Telugu fiction of that time (1960s-80s). Both the authors manually annotated all the sentences, and disagreements were adjudicated after discussion. Annotation was done using the Brat (Stenetorp et al., 2012) tool and the conversion to CONLL format was done using a Python script. We annotated the sentences with UD POS tags and dependency relations (annotation guidelines available on the UD website). The whole process of annotation and correction process took 4 months. The following sections outline our annotation decisions with examples.

---

2 Available at http://kcis.iiit.ac.in/
3 The Dravidian language family is one of the four language families spoken in India (the others being Indo-European, Austro-Asiatic, and Sino-Tibetan).
4 The closest parallel of a negative verb is the negative auxiliary verb in Finnish http://wals.info/chapter/112 and in Kurmanji (tune).
5 http://universaldependencies.org/guidelines.html
5 Part-of-speech annotation

UD specification defines 17 POS tags, of which we used 14 tags in Telugu. We did not use the tags - X, SYM and AUX. X was not used as there were no instances of unanalyzable foreign words that are not loan words in the corpus. SYM was not used as there were no symbols in the sentences. Words that resemble auxiliary verbs in Telugu also function as main verbs in the sentence. Hence, we did not use the AUX tag in our annotations. While we largely followed UD guidelines for POS tagging, we also made accommodations for some language specific phenomena. Among the open-class words, while Nouns, Proper Nouns and Verbs are relatively straightforward to tag, we made specific annotation decisions for adjectives and adverbs. Verbs functioning as Nouns or Adjectives were tagged Verbs but annotated with appropriate dependency relation (e.g., acl for verbal adjectives) to the head. We mark the morphological feature VerbForm with Vnoun and Part respectively.

Adjectives (ADJ): Adjectives in Telugu are indeclinable. Oblique nouns functioning as adjectives are tagged as NOUN with the relation nmod:poss to the head noun. Adjectives with a pronominal suffix (e.g., manci=vaː=ã u good-3-SG-M. suffix, meaning: good one) are tagged as PRON and not adjective, as they refer to entities. Figure 1a illustrates a sentence with pronominalized adjective in predicate position with words transcribed in IPA. Adjectives denoting dimensions such as tall (poː̯uɡu) or short (poː̯uí) do not need a pronominal suffix to function as the root of a sentence non-verbal construction. In such cases, we treat such as adjectives as abstract nouns and mark the POS tag as NOUN.

Adverbs (ADV): Krishnamurti and Gwynn (1985) and Krishnamurti (2003) note that adverbs of time and place behave as nouns (can inflect with case markers) in Telugu. We adopt the judgment into our treebank and mark all adverbs of time and place as nouns. We annotate an inflected time or location noun as NOUN and annotate it as the dependent of the dependency relationsobl or obl:tmod. This is shown in Figures 1b and 1c.

Adpositions (ADP): Telugu uses postpositions and suffixes to denote cases. Postpositions are tagged as ADP and are dependent of nominals through case relation. Some adverbs indicating temporal or location information that appear as nominal modifiers are also tagged ADP.

Determiners (DET): UD guidelines distinguish 6 kinds of determiners. Of those, Telugu does not have articles and possessive determiners. We mark distal/proximal demonstratives and interrogative determiners that precede a nominal as DET. Telugu does not have relative pronouns and forms relative clauses through nominalization or verbal adjectives.

Numerals (NUM): Following UD guidelines, we tagged all numbers, fractions and multi-word numeric expressions with the NUM tag. However, numbers can also function as adjective, adverb or noun in Telugu, and can be inflected. Inflected numbers which do not appear in a multi-word numeric expression...
are marked according to their syntactic function. UD guidelines also describe the tagging of non-cardinal numbers according to their syntactic function.6

6 Morphology

Telugu verbs show agreement with the agent in number, gender and person. Telugu has two genders: masculine and non-masculine and we will annotate with the same categories. Telugu has two numbers: singular and plural. Telugu nominals show highly inflected case system with nominative, dative, instrumental, genitive, commitative, ablative, and locative cases. Postpositions also function as adessive (miđa “on”), purposive and comparative cases. Complex cases are formed through a combination of base case markers and postpositions.

Telugu verbs show tense, aspect, and mood. Verbs are typically active voice and passive constructions are not common. Causative constructions (Voice=Cau) are formed by adding -inc to the transitive verb. Telugu also has a reflexive suffix -kon that is added to causative and transitive verb stems to denote that the agent is also the patient. We mark such a reflexive verb with Reflex=Yes. There are two tenses: past and non-past. Telugu does not have a negative particle and shows negation through -a- marker that occurs before index markers. Verb can show aspect: habitual (Hab), progressive (Prog), perfect (Perf), prospective (Prosp) which are available in UD. The mood features are imperative (Imp), conditional (Cnd), potential (Pot), necessitative (Nec), inceptive (Inc), hortative (Hor).7 Morphological annotation is not a part of the UD 2.1 release for Telugu and is a part of future work.

7 Universal Dependency Relations

Our treebank has 42 dependency relations, of which 11 are language specific. They are listed below in Table 1. Relations that are not seen in other language UD treebanks are marked with *. 

<table>
<thead>
<tr>
<th>Relation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl:relcl</td>
<td>Relative clause</td>
</tr>
<tr>
<td>advcl:cond</td>
<td>Conditional Adverbial clause</td>
</tr>
<tr>
<td>compound:lvc</td>
<td>Light verb construction</td>
</tr>
<tr>
<td>compound:redup</td>
<td>Reduplicative construction</td>
</tr>
<tr>
<td>compound:svc</td>
<td>Serial verb construction</td>
</tr>
<tr>
<td>nmod:cmp*</td>
<td>Nominal comparative modifier</td>
</tr>
<tr>
<td>nmod:poss</td>
<td>Nominal possessive modifier</td>
</tr>
<tr>
<td>nmod:tmod</td>
<td>Nominal temporal modifier</td>
</tr>
<tr>
<td>nsubj:inc</td>
<td>non-canonical subjects (e.g., dative subjects)</td>
</tr>
<tr>
<td>obl:tmod</td>
<td>Oblique case-temporal</td>
</tr>
<tr>
<td>obl:cau*</td>
<td>Oblique case-causative (Section 6)</td>
</tr>
</tbody>
</table>

Table 1: Language Specific Dependency Relations for Telugu

While some relations such as acl:relcl and nmod:poss exist in several other language UD treebanks, other relations are not very common. Some of them are discussed below.

Light verb constructions: Light verbs are noun-verb constructions where the semantic content is in the noun even if the syntactic head is the verb. These constructions are wide spread in Hindi-Urdu (Butt, 2010; Vaidya et al., 2016). However, the Hindi UD treebank (converted from Paninian dependency treebank) does not seem to tag this construction specifically though it is annotated with pof (part-of) relation in the original Paninian treebank. We tag this construction explicitly using compound:lvc. In UD 2.1, this construction is explicitly marked in Farsi, Kazakh, Kurmanji, Marathi, Turkish and Uyghur along with Telugu. Recent work in Hungarian (Vincze et al., 2017) described these constructions using the label dobj:lvc. Light verb construction is illustrated in figure 2 where a noun (start) followed by a verb is used as a light verb compound.

6http://universaldependencies.org/u/pos/NUM.html
7Inceptive and Hortative moods are not available in UD.
Kamala started singing a song.

**Reduplication:** Reduplication is the morphological process in which whole or parts of the word are repeated to denote a syntactic function. Reduplication (both partial and complete) is a common phenomenon in several languages although it is explicitly marked in only five other UD 2.1 languages – Hindi, Kurmanji, Marathi, Turkish and Uyghur – with the relation `compound:redup`. We mark all reduplicated words with this relation and treat the final word as the head. In Telugu, reduplication can occur across POS categories such as determiner, verbs, adjectives, nouns, and adverbs. We show examples of verb and adjective reduplication in figure 3.

**Serial Verbs:** The Dravidian comparative literature defines serial verbs as a series of finite verbs which are present in Old Telugu but absent in Modern Telugu. There is no limit to the number of participating verbs in such a construction. We employ the definition of serial verbs from Velupillai (2012, 332) that a series of verbs referring to a single event is labeled as serial verb. These are different from other compound verbs such as V-V complex predicates in that they describe a sequence of actions. We mark such constructions with `compound:svc` (cf. Figure 4). The Dravidian comparative literature treats these constructions as adverbial clauses.

**Non-verbal predication:** In this paragraph, we present non-overt copula (cf. Figure 5a) and negative verb (cf. Figure 5b) which is specific to Telugu. Equative, attributive, possession, and benefaction constructions consist of NP+NP and lack an overt copula. Location construction (negation variant) shown in Figure 5b shows agreement and does not fall under the definition of non-verbal predication in UD.

**Genitives:** Genitive constructions can be formed through a preceding nominal dependent in nominative case, postposition (`yokka`), and oblique noun. We mark all these relations as `nmod:poss` (cf. figure 6).

**Comparatives:** Comparative constructions are formed through a special postposition `kanṭe` which is marked as a dependent of the second nominal through `case` relation (cf. Figure 7).  

---

*At the time of submission of the paper, we marked the relation between the two nominals using `nmod:comp` relation. We mark the relation between the second nominal and the root noun with `obl` relation.* We thank one of the reviewers for pointing...
I went to the library yesterday.

Figure 4: Serial verb construction

She is my mother. She was(is) not in the room.

(a) Non-overt copula (b) Negative existential verb

Figure 5: Non-verbal predication in Telugu.

Figure 6: Genitive formation strategies

Ramu is taller than Kamala.

Figure 7: Comparative construction without an overt copula

**Dative subjects:** Typically, NPs that occur at the sentence-initial position are in nominative case (unmarked). However, stative verbs such as “to know” and intransitive verbs such as “to want” do not show any agreement with any of the NPs in the sentence. In such cases, the NP in initial position is marked with dative case (Sridhar, 1979; Nizar, 2010). We mark the syntactic relation between the verb and the dative NP with nsubj:nc. Although dative NP occurs in sentence initial position, the free word order allows the dative NP to be moved to a non-final position in the sentence. A dative NP (annotated as

this mistake.
Adverbial clauses: Telugu forms adverbial clauses through converbs. The final verb in the sentence is a finite verb which is treated as the root of the sentence. The subject of the embedded clause can be co-referential (cf. Figure 8) when the non-finite verb is marked for perfective or progressive aspect. In such a case, we annotate the subject to be the dependent of the main verb. Subjects of the main and subordinate clauses cannot be co-referential when the non-finite verb is marked for conditional or concessive moods.

Relative clauses: There is no relative pronoun in Telugu and relative clauses are formed through verbal adjectives. There are no expletive nominals in Telugu and cleft constructions are formed through pronominalized verbal adjectives. We analyze cleft sentences as relative clauses (cf. figure 9).

Nominalized clauses: Non-finite verbs are nominalized by adding -atam. A nominalized verb can then be the head of a subordinate clause which can be the subject or object of the main verb (cf. figure 2). We annotate a nominalized verb clause as csubj (functions as subject) and ccomp (functions as object), respectively.

8 Tagging and Parsing Experiments

As a demonstration of the usefulness of our treebank in real world settings, we evaluated POS tagging and parsing models trained using UDPipe (Straka et al., 2016). UDPipe is a free, open-source, and language agnostic pipeline for training and evaluating NLP models for lemmatization, POS tagging and dependency parsing.

We split our treebank into 80-10-10 for training, development, and testing; and trained POS tagging and parsing models. Both training and evaluation was performed with UDPipe-1.2 on a Linux machine.

---

9We follow the Persian UD annotation guidelines (Seraji et al., 2016) in this case and name the dependency relation as nsubj:nc.
We report POS tagging accuracy, Labeled Attachment Score (LAS), and Unlabeled Attachment Score (UAS) on test set (after parameter tuning on development set) in Table 2. We also trained and evaluated a second parsing model on gold POS tags and found that the LAS and UAS scores are better than the joint model for predicting POS tags and dependency relations. We expect the POS tagging results to be high since nouns and verbs make the bulk of the part-of-speech tags in Telugu.

<table>
<thead>
<tr>
<th>Input features</th>
<th>POS Acc.</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagging + Parsing</td>
<td>90.43%</td>
<td>74.76%</td>
<td>87.79%</td>
</tr>
<tr>
<td>Parsing (Gold POS tags)</td>
<td>–</td>
<td>78.50%</td>
<td>89.74%</td>
</tr>
</tbody>
</table>

Table 2: Preliminary tagging and parsing results with UDPipe.

Previous work on Telugu dependency parsing – trained and evaluated with Paninian dependency labels – report the highest LAS of 70.15% (Husain et al., 2010) and best UAS of 90.5% (Kanneganti et al., 2016). While our LAS results are higher than both the previous results, the UAS results are slightly lower; however, a direct comparison is not possible due to the unavailability of the training data for these results and also due to different annotation schemes.

9 Conclusion

In this paper, we presented the first publicly available treebank for Telugu annotated in Universal Dependencies framework. We annotated POS tags and dependency relations from scratch for 1328 sentences. We trained and evaluated two parser models using UDPipe on the training split of the treebank and found that the parser performs within the range reported in the previous experiments.

As a part of future work, we intend to add morphological annotations to the treebank. It would also be interesting to compare different parsers on this treebank data. We are currently working towards expanding the treebank to include at least 100,000 tokens from Telugu Wikipedia. We plan to achieve this in a semi-automated fashion by running a trained parser model on Wikipedia sentences and then, manually checking and correcting for errors. We are also in the process of augmenting the treebank with fine-grained POS tags designed for Indian languages (Choudhary and Jha, 2011). The average sentence length in our treebank corpus is rather small (~ 5 tokens per sentence) whereas Wikipedia sentences are typically much longer. We intend to analyze how accurate can an automatic parser trained on grammar book examples would be when faced with longer sentences, with possibly complicated syntactic structures.

Acknowledgments

The first author was supported by ERC Advanced Grant 324246 EVOLAEmp and a NRC LightHouse grant which is gratefully acknowledged. We thank Çağrı Çöltekin, Lilja Øvrelid, and Viswanath Naidu for all the discussions and input which proved useful in the development of the treebank. Finally, we thank the anonymous reviewers for the comments that helped improve the paper.

References


