A Fully Expanded Dependency Treebank for Telugu

Sneha Nallani, Manish Shrivastava, Dipti Misra Sharma
International Institute of Information Technology, Hyderabad

Introduction

- The available Paninian dependency treebank(s) for Telugu is annotated only with inter-chunk dependency relations.
- In this paper, we automatically annotate the intra-chunk dependencies in the treebank.
- Annotating intra-chunk dependencies leads to a complete parse tree for every sentence in the treebank.
- Having complete parse trees is essential for building robust end to end dependency parsers, making use of readily available parsers.
- We propose a few additional intra-chunk dependency relations for Telugu.
- We also convert the treebank annotated with Anncorra part-of-speech tagset to the latest BIS tagset.
- The final treebank is made publicly available.

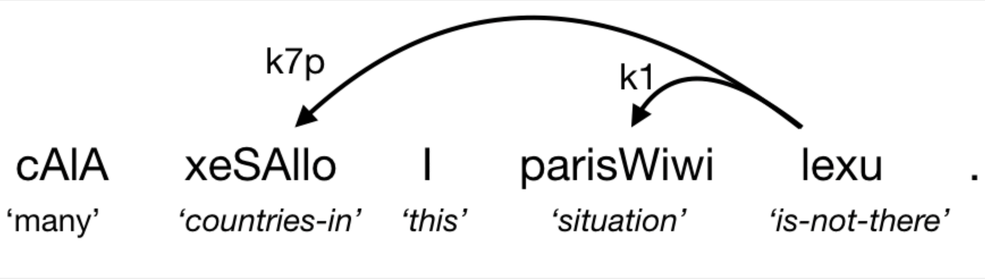
Telugu Treebank

- IIIT-H Telugu treebank with 1600 sentences is made available in ICON 2009 tools contest.
- Combined with HCU Telugu treebank containing approximately 2000 sentences.
- Clean up the treebank by removing sentences with wrong format or incomplete parse trees etc.
- Treebank annotated at inter-chunk level in Shakti-Standard Format (SSF)

No. of sentences	3222
Average sentence length	5.5 words
Average no. of chunks in sentence	4.2
Average length of a chunk	1.3 words

```
<Sentence id='10'>
1     ((     NP    <fs af='xeSaM,n,,pl,,,lo,lo' head="xeSAllo" drel='k7p:VGF' name='NP'>
1.1   cAlA QT_QTF   <fs af='cAlA,avy,,,,,0,0_avy'>
1.2   xeSAllo    N_NN       <fs af='xeSaM,n,,pl,,,lo,lo' name="xeSAllo">
      ))
2     ((     NP    <fs af='parisWiwi,n,,sg,,d,0,0' head="parisWiwi" drel='k1:VGF' name='NP2'>
2.1   I      DM_DMD <fs af='I,avy,,,,,,' poslcat="NM">
2.2   parisWiwi  N_NN       <fs af='parisWiwi,n,,sg,,d,0,0' name="parisWiwi">
      ))
3     ((     VGF <fs af='gala,v,fn,sg,3,,a,a' head="lexu" name='VGF'>
3.1   lexu  V_VM       <fs af='gala,v,fn,sg,3,,a,a' name="lexu">
3.2   .      RD_PUNC       <fs af='.,punc,,,,,,' poslcat="NM">
      ))
</Sentence>
```

Inter-chunk dependency annotation in SSF format



Inter-chunk dependency tree

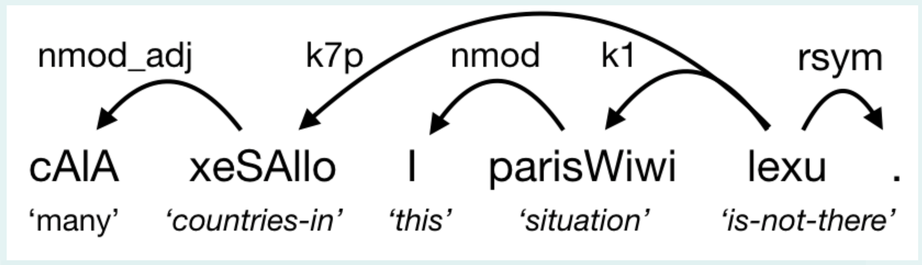
Anncorra to BIS POS conversion

- **Anncorra** tagset was developed as part of ILMT project and consists of 26 tags.
- **BIS** is a hierarchical tagset being developed as a unified POS Standard in Indian Languages.
- We annotate with the most fine grained BIS tag and fall back to the parent tag if finer tag can't be determined.
- In Anncorra schema verb finiteness is marked at chunk level and in BIS, at word level.
- Other Anncorra tags diverging into finer BIS tags are for function words. Lists of words belonging to finer BIS tags are created and used for annotation.

Anncorra POS tag	BIS POS tag
PRP (Pronoun)	PR_PRP, PR_PRF, PR_PRL, PR_PRC, PR_PRQ
DEM (Demonstrative)	DM_DMD, DM_DMR, DM_DMQ
VM (Main verb)	V_VM_VF, V_VM_VNF, V_VM_VINF, V_VM_VNG, N_NNV
CC (Conjunct)	CC_CCD, CC_CCS
WQ (Question word)	DM_DMQ, PR_PRQ
SYM (Symbol)	RD_SYM, RD_PUNC
RDP (Reduplicative)	-
*C (Compound)	-

Fine grained BIS tags corresponding to Anncorra tags.

Intra-chunk dependency annotation



Intra-chunk dependency tree.

- Kosaraju et al. (2012) proposed 12 intra-chunk dependency labels and guidelines for annotating intra-chunk dependencies in SSF format for Hindi.
- Bhat (2017) propose to annotate intra-chunk dependencies for Hindi and Urdu using a shift-reduce parser and Context Free Grammar(CFG) rules.
- We follow Bhat (2017) approach and write the CFG for Telugu and propose 4 additional intra-chunk dependencies for Telugu.

nmod__adj	adjectives modifying nouns or pronouns
lwg__psp	post-positions
lwg__neg	negation
lwg__vaux	verb auxiliaries
lwg__rp	particles
lwg__uh	interjection
lwg__cont	continuation
pof__redup	reduplication
pof__cn	compound nouns
pof__cv	compound verbs
rsym	symbols
nmod__wq *	question words modifying nouns
nmod *	proper nouns, pronouns etc modify a noun or pronoun
intf *	intensifier modifying adjectives, adverbs
adv *	adverbs

Intra-chunk dependency labels. The ones marked with * are proposed for Telugu

Results

- We evaluate on a test set of 106 sentences.

Test sentences	LAS	UAS
106	93.7	95.8

- Almost all of the wrongly annotated chunks are because of POS errors or chunk boundary errors.

Conclusion

- We automatically annotate the Telugu dependency treebank with intra-chunk dependency relations thus finally providing complete parse trees for every sentence in the treebank.
- We also convert the Telugu treebank from AnnCorra part-of-speech tagset to the latest BIS tagset.
- We make the fully expanded Telugu treebank publicly available to facilitate further research.

sneha.nallani@research.iiit.ac.in m.shrivastava@iiit.ac.in dipti@iiit.ac.in