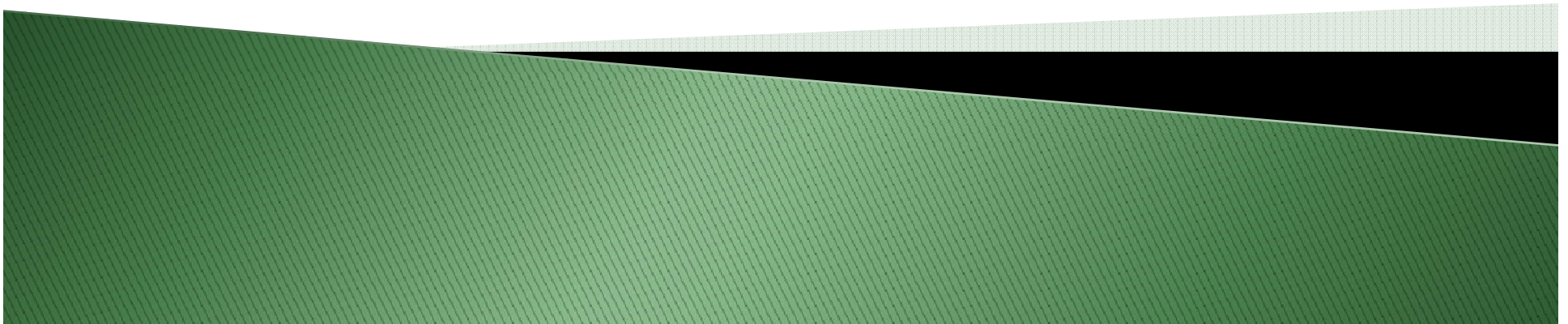


# On the Role of Morphosyntactic Features in Hindi Dependency Parsing

Bharat Ram Ambati\*, Samar Husain\*, Joakim Nivre† and Rajeev Sangal\*

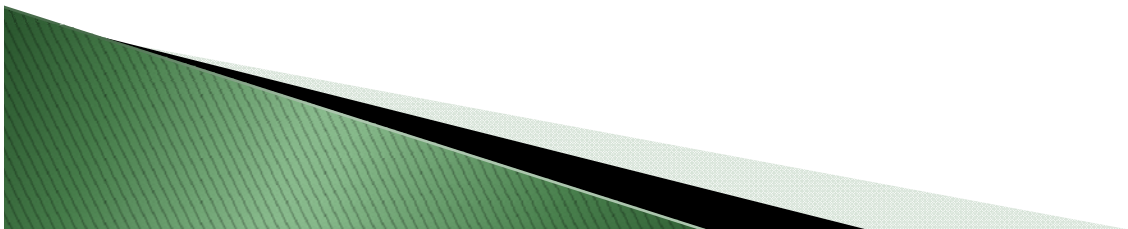
\*Language Technologies Research Centre, IIT–Hyderabad, India.

†Department of Linguistics and Philology, Uppsala University, Sweden.



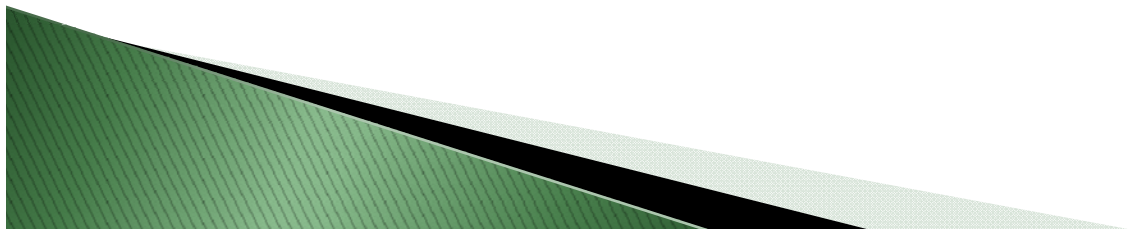
# Hindi Dependency Parsing

- ▶ Hindi is a free word order language with relatively rich morphology
- ▶ Dependency treebank (Begum et al., 2008) based on the computational Paninian grammar (Bharati et al., 1995)
  - *karaka*: Syntactico–semantic labels
- ▶ In this work we use MaltParser (Nivre, 2008) to improve the performance of Ambati et al. (2009) and Nivre (2009) on the ICON09 dependency parsing dataset (Husain, 2009)
  - Training:1500, Devel:150, Test:150
  - Avg. sentence length: 19.85



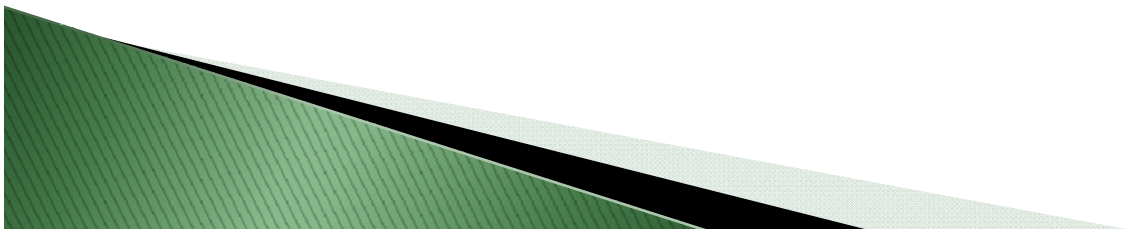
# Feature pool (Experiments 1–10)

	PTAG	CTAG	FORM	LEMMA	DEPREL	CTAM	OTHERS
Stack: <i>top</i>	1	5	1	7		9	
Input: <i>next</i>	1	5	1	7		9	
Input: <i>next+1</i>	2	5	6	7			
Input: <i>next+2</i>	2						
Input: <i>next+3</i>	2						
Stack: <i>top-1</i>	3						
String: predecessor of <i>top</i>	3						
Tree: head of <i>top</i>	4						
Tree: leftmost dep of <i>next</i>	4	5	6				
Tree: rightmost dep of <i>top</i>					8		
Tree: left sibling of rightmost dep of <i>top</i>					8		
Merge: PTAG of <i>top</i> and <i>next</i>							10
Merge: CTAM and DEPREL of <i>top</i>							10

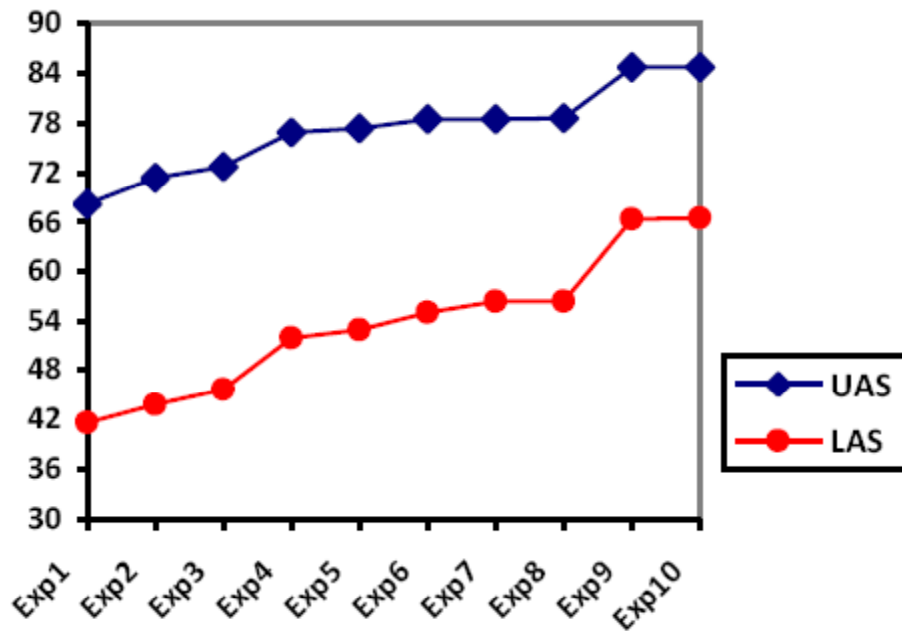


# Experiment 1–10

- ▶ 5-fold cross validation on training + devel data
- ▶ Experiment 1: PTAG and FORM
- ▶ Experiment 4: PTAG of nodes in the partially built tree
- ▶ Experiment 5–7: CTAG, FORM, and LEMMA
- ▶ Experiment 8: DEPREL of nodes in the partially formed tree
- ▶ Experiment 9: CTAM attribute of *top and next*
- ▶ Experiment 10: Conjoined features: POS of next and top and of CTAM and DEPREL of top



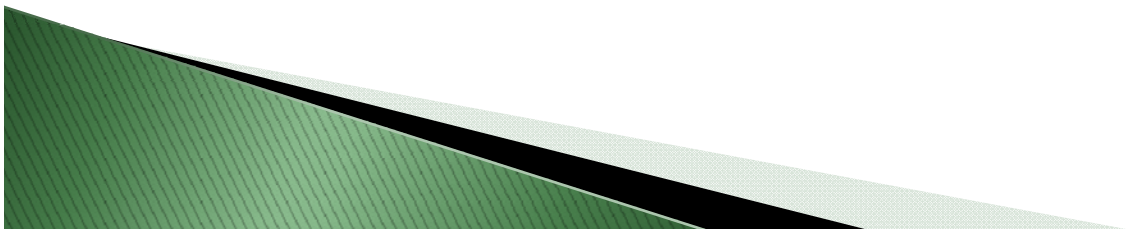
# Performance



- ▶ CTAM gave the greatest improvement ~10% jump in both LAS and UAS.
- ▶ It consists of two morphosyntactic features: case markers (as suffixes or postpositions) and TAM markers.
- ▶ It helps because
  - case markers are important surface cues that help identify various dependency relations, and
  - there exists a direct mapping between many TAM labels and the nominal case markers because TAMs control the case markers of some nominals.

# Results

<b>System</b>	<b>LAS</b>	<b>UAS</b>
Ambati et al. (2009a)	74.5	90.1
Nivre (2009b)	73.4	89.8
Our system	76.5	91.1



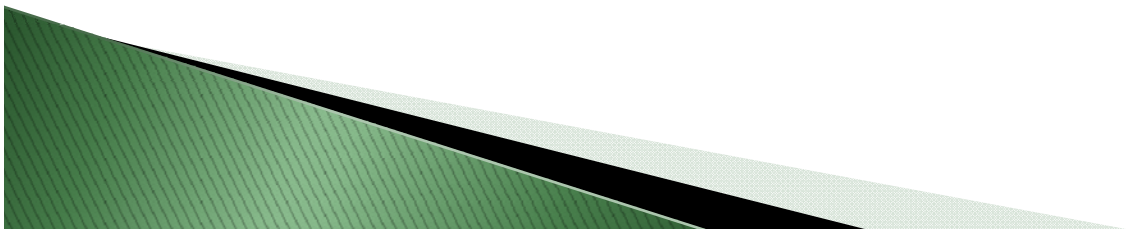
# Error Analysis

- ▶ Simple Sentences
  - the attachments are mostly correct, the dependency labels are error prone
- ▶ Embedded Clauses
  - argument sharing, ambiguous attachment site in participles
- ▶ Coordination
  - Long distance dependencies
  - Varying behavior, lack of morph features
- ▶ Complex predicate
  - Confusion with other labels
- ▶ Non-projectivity
  - ~14% non-projective arcs (Mannem et al., 2009)
  - Majority inter-clausal
  - 0/11 identified in test
- ▶ Long distance dependencies



# References

- ▶ B. R. Ambati, P. Gadde, and K. Jindal. 2009. Experiments in Indian Language Dependency Parsing. *Proc. of ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, 32–37.
- ▶ R. Begum, S. Husain, A. Dhvaj, D. Sharma, L. Bai, and R. Sangal. 2008. Dependency annotation scheme for Indian languages. *Proc. of IJCNLP*.
- ▶ A. Bharati, V. Chaitanya and R. Sangal. 1995. *Natural Language Processing: A Paninian Perspective*, Prentice–Hall of India, New Delhi.
- ▶ S. Husain. 2009. Dependency Parsers for Indian Languages. *Proc. of ICON09 NLP Tools Contest: Indian Language Dependency Parsing*.
- ▶ P. Mannem, H. Chaudhry, and A. Bharati. 2009a. In–sights into non–projectivity in Hindi. *Proc. of ACL–IJCNLP Student Research Workshop*.
- ▶ J. Nivre. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics* 34(4), 513–553.
- ▶ J. Nivre. 2009. Parsing Indian Languages with Malt–Parser. *Proc. of ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, 12–18.



Questions??

