GEORGETOWN UNIVERSITY

MASALA: Modelling and Analysing the Semantics of Adpositions in Linguistic Annotation of Hindi

Aryaman Arora Nitin Venkateswaran Nathan Schneider
{aa2190,nv214,nathan.schneider}@georgetown.edu

nert.georgetown.edu

Introduction

- **Hindi** (and Urdu) has a complex inventory of case markers and postpositions for marking semantic relations between words
  - *Case markers*: small set of markers that indicate core arguments to a verb, other basic relations
  - *Focus markers*: discourse relations, emphasis
  - *Postpositions*: large class of markers for narrow semantic relations, verbal adjuncts
- Understanding these relations is a difficult task for NLP
- Potential upstream benefits: *semantic role labelling*, *translation*
- We created a Hindi corpus annotated with coarse semantic labels from the **SNACS** formalism and attempted automatic labelling with language models!

SNACS

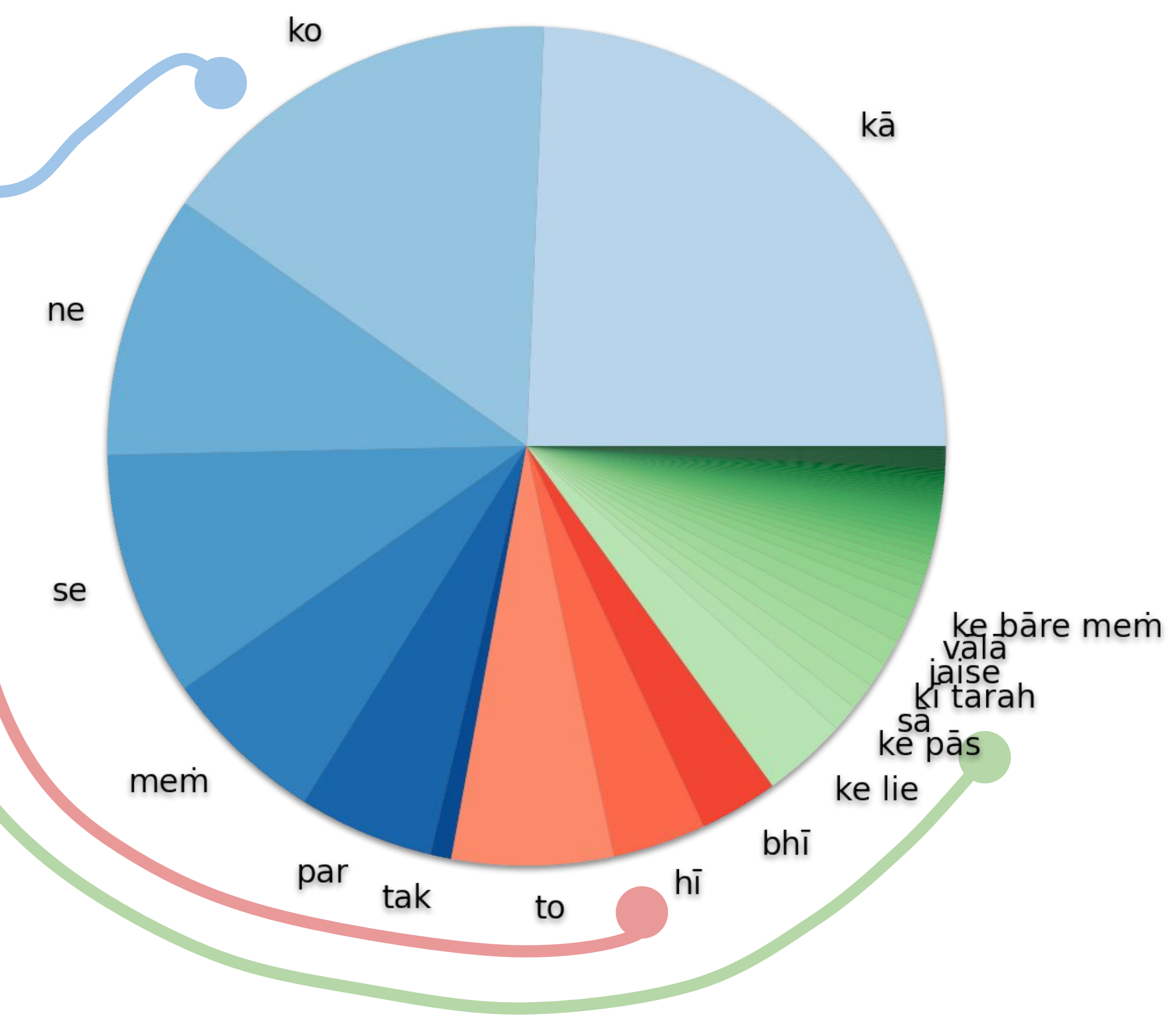
- SNACS is the **Semantic Network of Adposition and Case Supersenses**, already applied on English (L1 and L2), Korean, Mandarin, and German (Schneider et al., 2018a, 2020)
- Related to linguistic theories of argument structure and theta roles: Agent, Theme, Recipient, Causer, ...
- Construal system: attempts to separate syntax from semantics
  - Experiencer⇝Agent: a subject marked with the ergative (strongly agentive) with an experiencer predicate (e.g. feel)
  - Scene Role⇝Function
- **Linguistic issues in annotating Hindi**
  - Syntactic function of some case markers is hard to label
  - Non-nominative/ergative subjects
  - Causative constructions: are animate Instruments a thing?
  - Emphatic particles

(3) vah g^har **ke_pās**$_{\text{LOCUS}}$ hai
3SG home near COP.IND.3SG
'He is near the house.'

(4) maiṁ us **ko**$_{\text{THEME}}$ k^hā-tā hūṁ
1SG 3SG ACC eat-IPFV.M.SG COP.IND.1SG
'I eat that.'

(5) maiṁ **ne**$_{\text{EXPERIENCER⇝AGENT}}$ nadī
1SG ERG river
ke_pār$_{\text{LOCUS⇝PATH}}$ ek baccā dekh-ā
across one child.NOM see-PFV.M.SG
'I saw a child across the river.'

Figure 1: **Target types in the corpus**



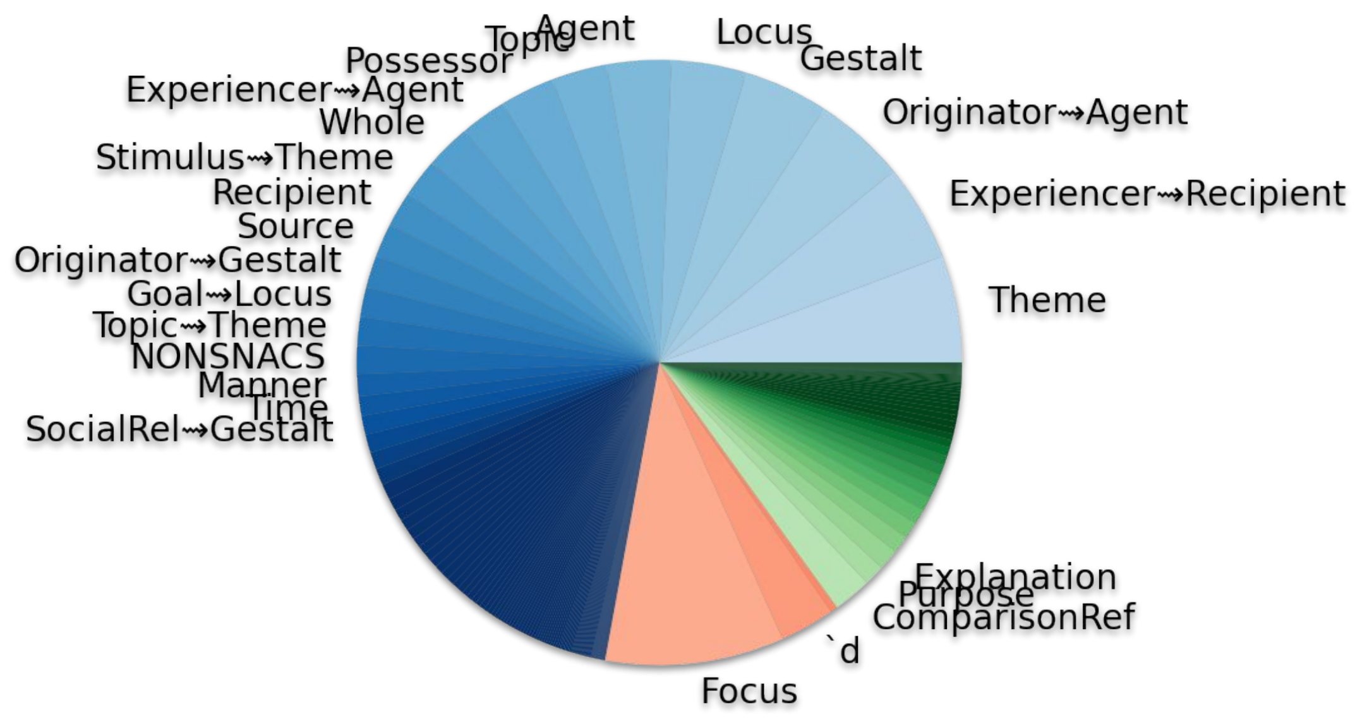


Figure 2: **Construals**

Corpus analysis

- Breakdown of target and annotation types shown above
- **Agreement**: Cohen's κ for doubly-annotated targets was 0.78 on scene roles, 0.85 on functions, and 0.73 on construals (both together)
  - Apparently syntax is easier to categorise than semantics
- **Semantic diversity**: We estimated the entropy of the distribution over scene roles for each token type in the corpus. Found that case markers have very high entropy: highly semantically diverse.

Automatic tagging

- We extended the **lexical semantic recognition** (Liu et al., 2021) task to Hindi: automatic tagging of coarse supersenses on case/adpositions
- Data processing
  - Convert labels to BIO-tagging scheme
  - Split into 80/10/10 train/dev/test set, check performance improvements on dev to stop training
- *Architecture*: contextual language model → biLSTM → CRF to output tags
  - We tested various BERT-like masked language models for Hindi, with a 2-layer biLSTM with dropout of 0.3, then to a CRF
  - *Hyperparameters*: {30, 60} epochs, {0.0001, 0.0002, 0.0005, 0.001} learning rate, {64, 128, 256, 512} LSTM layer size)
  - Found to be better to use biLSTM+CRF than just biLSTM or Transformers
- *Results*: IndicTransformers BERT is the best, with comparable numbers to past work on English (two on left below)

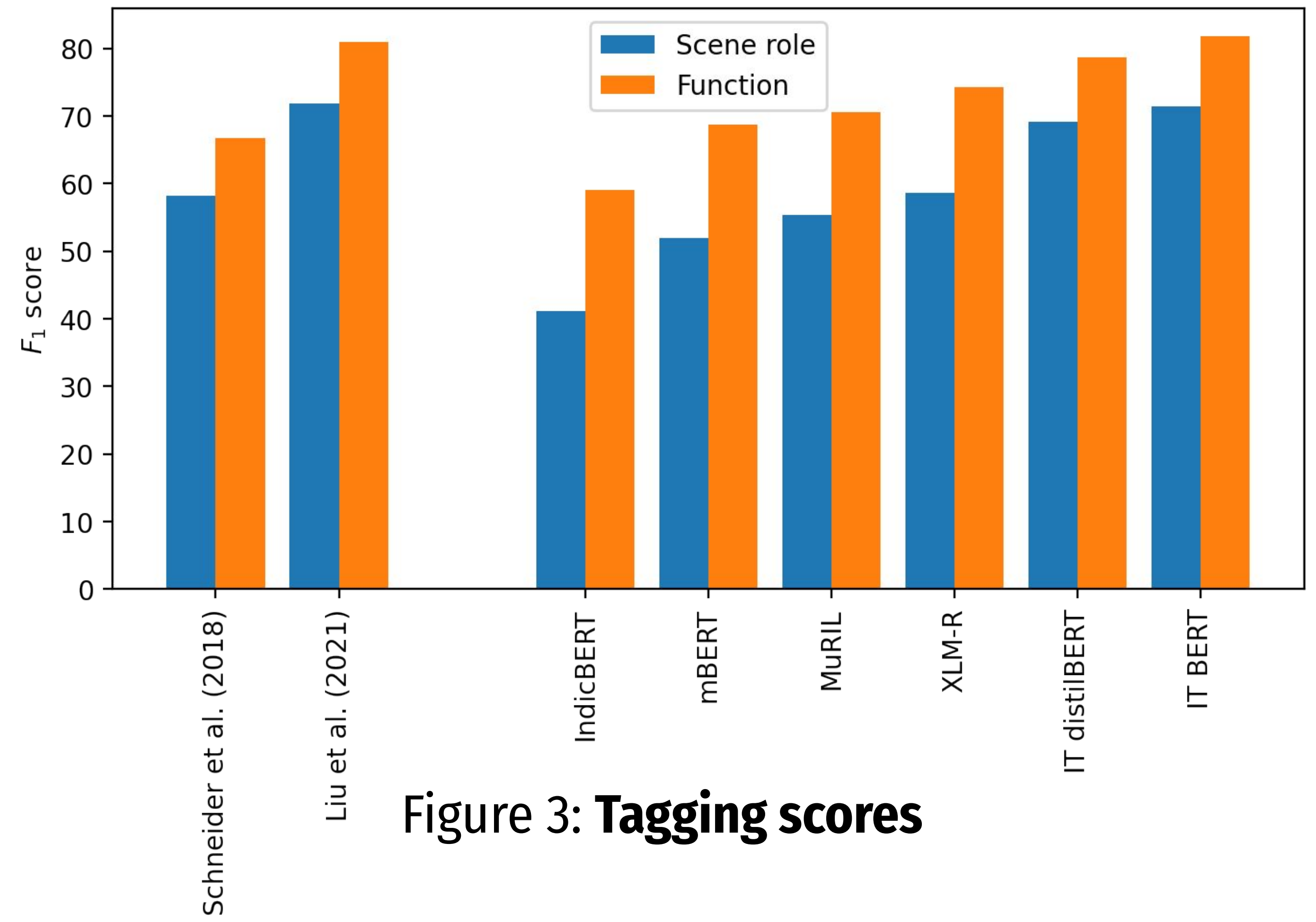


Figure 3: **Tagging scores**

References

- Liu, Nelson F., Hershcovich, Daniel, Kranzlein, Michael, and Schneider, Nathan (2021). Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021).*
- Schneider, Nathan, Hwang, Jena D., Srikumar, Vivek, Prange, Jakob, Blodgett, Austin, Moeller, Sarah R., Stern, Aviram, Bitan, Adi, and Abend, Omri (2018a). Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL.*
- Schneider, Nathan, Hwang, Jena D., Bhatia, Archna, Srikumar, Vivek, Han, Na-Rae, O'Gorman, Tim, Moeller, Sarah R., Abend, Omri, Shalev, Adi, Blodgett, Austin, and Prange, Jakob (2020). Adposition and Case Supersenses v2.5: Guidelines for English. *arXiv:1704.02134 [cs].*