# What Do Hindi Native Speakers Predict?

## INVESTIGATING VERB CLASS, VERB MORPHOLOGY AND WORD ORDER

**Apurva and Samar Husain**
Indian Institute of Technology, Delhi

# Introduction

- *abhay ne roti .....*

  Abhay ERG bread ….

- It has been established that human sentence processing involves prediction, i.e. we do not wait for the entire utterance to complete before making sense of the utterance
- Prediction can be thought as some kind of top-down process wherein the upcoming linguistic material becomes activated before it is actually encountered

# Introduction

- While it is undisputed that we predict, the exact mechanism of the prediction system is still unclear

- It is also unclear as how the prediction system interacts with other cognitive processes/resources (e.g. working memory)

# Broad goal of this project

- What are the processes underlying moment-by-moment predictions of a Hindi native speaker?
- How do these interact with working memory constraints?
- What does it tell us about human sentence processing in general?

# More specifically

- We investigate:
  - verb class,
  - word order,
  - verb morphology

- In order to investigate these, we vary:
  - the number of (animate) noun phrases,
  - the case markers on the noun phrases

# Experiment

- We conducted a sentence completion study
  - Items were presented using Linger in self-paced reading paradigm
  - Participants read incomplete sentences and had to complete it meaningfully
- Items were randomized
- 30 Hindi native speakers from Jawaharlal Nehru University and IIT Delhi participated in the experiment

# Design

- No. of noun phrases: 1, 2, 3
  - *ladkaa*, *dost*, *mahilaa*
- Case markers: *0*, *ne*, *ko*, *se*
- There were:
  - 4 one noun phrase conditions
  - 16 two noun phrase conditions
  - 64 three noun phrase conditions
- There are 29 subjects, and 84 items
  - In all there were 2436 data points
    - Data from 1 subject was removed because he/she did not understand the task
    - 1 condition only had 6 data points due to a coding error

# Design

- 1 NP examples
  - *ladke ne*
  - *ladke ko*
  - ...
- 2 NP examples
  - *ladke ne dost ko*
  - *ladke ko dost se*
  - ...
- 3 NP example
  - *ladke ne dost ko mahilaa se*
  - ...
- So, all items were of the form
  - *ladka/ladke.ne/ko/se dost.0/ne/ko/se mahilaa.0/ne/ko/se*

# Coding

- After running the experiment, data transcription involved coding the following information based on participant response
  - **Grammaticality**: 1 (grammatical), 0 (ungrammatical)
  - **Word Order**: 1 (canonical), 0 (non-canonical)
  - **Verb class**: T (transitive), IN (intransitive), DT (ditransitive), EXP (experiencer), COP (copula), CAUS (Causative)
  - **Verb morphology**: Perf (perfective), Obl (obligational), …
  - …
  - …
  - … (some others that are not discussed here)

# Caveats

- All sentences presented in isolation
  - It is known that discourse will have some role to play in predicting word order
- Broad verb classes
  - Some distinctions currently not made
- Only animate nouns used
- Current analysis of verb class/morphology only considers simple completions
  - e.g. completions with non-finite verbs are deemed complex and not considered
- Passives, argument ellipsis, imperatives, noun compounding not considered
- **These issues will be addressed in the future analysis/investigations**

# Grammaticality

- If prediction is pervasive during processing, then we expect that the noun phrases and case-markers will be successfully employed to produce grammatical continuations

- So, if a condition that can otherwise be completed grammatically gets high ungrammatical continuations, we need to understand why that is

# Grammaticality

- Very few ungrammatical cases in 1 NP conditions. Not surprising!
- Examples where ~50% instances are ungrammatical

| **2 NP** | **3 NP** |
|---|---|
| *ne, ne* | *ko, se, se* |
| *ko, ko* | *se, ne, se* |
| *se, se* | *se, ko, ko* |
| | *se, ko, se* |
| | *se, se, ko* |

# Grammaticality

- Other 3 NP cases such as (*ne*, *ko*, *se*), (*se*, *ko*, *ne*), etc. are grammatically completed
- Current pattern suggests that the main reason for ungrammaticality (in cases were grammatical continuations are possible) is some kind of cue-based interference. The interference happens because of the presence of duplicate case markings, e.g. (*se*, *ne*, *se*), (*se*, *ko*, *ko*), etc.

- **If this is true, then it would mean that certain valid case-marker combinations are not always beneficial for predicting the upcoming verb**

# Word Order

- Hindi is known to be a relatively free word order language
- Previous work on word order variation (e.g. Gambhir, 1981; Butt and King, 1996; Kidwai, 2000) are motivated by information structure and discourse considerations
- In this study, we only look at sentences in isolation

# Word Order

- In this study, 3.5% of the data has continuations with non-canonical word order
- What are these conditions?

| | |
|---|---|
| *ko, 0* | *se, ko* |
| *ko, 0, 0* | *se, 0, ko* |
| *ko, se, 0* | *se, ko, 0* |
| *se, 0* | *se, 0, 0* |

- These conditions have NP1 with a *ko* or *se* case marker
- **This means that *ko/se* is not preferred in the canonical subject position**
  - **Although they are grammatically permitted at this position**

# Word Order

- Interestingly, the order of case markers interact with word order prediction
  - *se* before *ko* in a 2 NP condition was not treated as a subject, rather *ko* was treated as the subject

# Verb class

- We know that case markers and no. of NPs together determine the verb class
- Not surprisingly we found evidence for this in our data

| | **Dominant verb class** |
|---|---|
| **1 NP** | Intransitive, Experiencer, Copula |
| **2 NP** | Transitive |
| **3 NP** | Causative |

# Verb class

- Similarly, case markers determine the verb class

| | 0 | ne |
|---|---|---|
| 1 NP | | |

- *ne* case marker insures the prediction of a transitive verb in close to 90% instances
- *ko* resulted in prediction of experiencer verbs (or verbs with dative subjects)
  - *ladke ko pyaar hai, ladke ko bukhar hai*
- *se* does not seem to affect verb class prediction as it is rarely treated as a subject

# Dynamic nature of verb class prediction

- How does the verb class prediction change as we get additional NPs?

**NP1.0 NP2.C***: Intransitive → Transitive
**NP1.0 NP2.C* NP3.se/ko**: Intransitive → Transitive → Causative

**NP1.ko NP2.se**: Transitive → Experiencer
**NP1.ko NP2.se N3.ne**: Transitive → Experiencer → Causative
**NP1.ko NP2.se N3.0**: Transitive → Experiencer → Transitive

**NP1.se NP2.se/ko**: Intransitive → Causative/Experiencer
**NP1.se NP2.ko NP3.ne**: Intransitive → Causative/Experiencer → Causative

# Dynamic nature of verb class prediction

- We can see that Hindi native speakers are constantly revising their predictions based on the new material
- We still need to investigate how this change interacts with word order
- Studying these changes will help us determine the processing cost of these factors through an online experiment

# Morphology

- We coded perfective, habitual, continuous, obligational, abilitative verb comple
  - Perfective~78%
  - Habitual~8%
  - Continuous~3%
  - Obligational+Abilitative~8%

- Data shows default completion is perfective

# Morphology

- Not surprisingly, *ne* triggers perfective morphology
- *se* marker resulted in the prediction of verbs with a specific (abilitative) verb complex
  - *ladke se padha nahi gya.*, *ladke se kaam nahi hua*
- conditions where we got habitual morphology (R data, subset HAB)
- conditions where we got continuous morphology(subset conditions for CONT)

- Data shows default completion is perfective

# Wrapping up

- Our analysis reiterates many established facts wrt the effect of no. of NPs and case marker on verb class and morphology
- However, we noted that the role of these factors is constrained, i.e. case markers need not always help in prediction
- In the absence of any prior discourse, non-canonical word order is rare. *se/ko* is not treated as a subject when it occurs in a sentence initial position
- Verb class prediction can change with additional material. These changes should lead to differing processing cost for different conditions

Thanks!