
Towards an Annotated Corpus of Discourse Relations in Hindi

Rashmi Prasad, Samar Husain, Dipti Misra Sharma and Aravind Joshi

Outline

- **Role of Annotated Corpora at the discourse level**
 - **Moving to annotations at the discourse level**
 - **A brief overview of the Penn Discourse Treebank (PDTB)**
 - **A brief overview of the Hindi Dependency Treebank**
 - **Hindi Discourse Treebank**
 - Subordinating Conjunctions
 - Subordinators
 - Coordinating Conjunctions
 - Discourse Adverbials
 - **Need to add the discourse layer over the dependency treebank**
 - Getting the correct argument structure
 - Lexical expression of discourse relations not shown explicitly
 - Semantics is underspecified
 - **Summary**
-

Role of Annotated Corpora at the Discourse Level

- Annotations at the discourse level
 - leading to certain levels of discourse processing, useful for applications
 - Moving from dependency annotation at the sentence level
 - to the annotation of discourse connectives and their arguments at the discourse level
-

Need for discourse relations

- Discourse relations provide a level of description that is
 - *theoretically interesting*, linking sentences (clauses) and discourse
 - *identifiable more or less reliably* on a sufficiently large scale
 - *capable of supporting a level of inference* potentially relevant to many NLP applications
-

The Penn Discourse Treebank (PDTB)

- Annotations of discourse relations holding between abstract objects (events, states and propositions, etc.) (PDTB Group, 2006)
 - Relations realized explicitly as **Explicit Connectives**
 - Implicit relations between adjacent sentences: **Implicit Connectives**
 - For both of these annotate the argument structure (connective and their two arguments)
 - In addition to the argument structure further aspects of the relation annotated:
 - Semantic classification of the connective (Explicit and Implicit)
 - Attribution of the connectives and their arguments (to capture the source and factuality)
-

PDTB

- **Explicit Connective (although, and, so, etc.)**
 - *The federal government suspended sales of U.S. savings bonds* because Congress hasn't lifted the ceiling on government debt.
- **Implicit Connective**
 - *Some have raised their cash positions to record levels.* Implicit=because (causal) High cash positions help buffer a fund when the market falls.

- Only 2 AO arguments, labeled *Arg1* and *Arg2*
- *Arg2*: clause with which connective is syntactically associated
- *Arg1*: the other argument

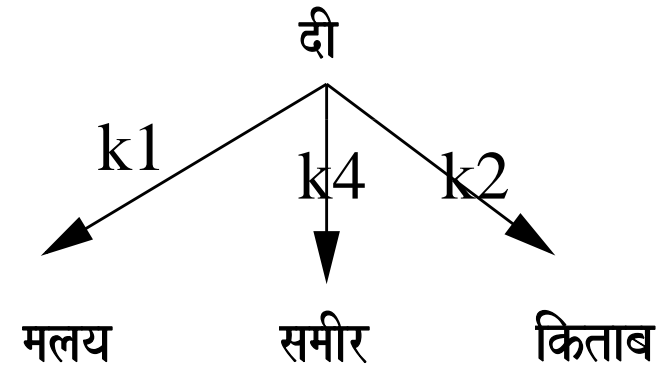
Hindi Dependency Treebank (LTRC, IIT-Hyd)

- Framework inspired by Panini's Grammar
 - Annotation over POS tagged and chunked corpus
 - syntactico-semantic relations
 - 28 labels
 - The labels provide us with a level of semantics (cf. PDTB)
 - Ongoing effort (Begum et al., 2008)
 - Present Status: 2500 sentences
 - Potential: ~ 200-300k words
 - Corpus
 - CIIL (Central Institute of Indian Languages)
 - Representative: Different fields such as newspaper text, literature, etc.
 - For both dependency annotation and discourse annotation
-

Hindi Dependency Treebank

मलय ने समीर को
किताब दी ।
malay ERG sameer DAT book
gave

“Malay gave the book to Sameer”



- k1: *karata karaka* (agent)
- k4: *sampradaan karaka* (beneficiary)
- k2: *karma karaka* (theme)

New Projects in Other Languages

- Chinese
- Czech
- Hindi
- Turkish

- Morphology may allow suffixes to serve as connectives
 - Possible overlap of a connective and an argument (arg2)
 - Possible divergence of semantics of the connectives across languages
 - ...
-

CONNECTIVE TYPE	HINDI	GLOSS	ENGLISH	NUM
Sub. Conj.	(क्यों)की..(इसलिए) (अगर यदी)..तब तो (जब).. तब तो जब तक.. तब तक (के लिए) जैसे ही..(तो) इतना एसा..की ताकी की	why-that..(this-for) (if)..(then) (when)..then when till..then till (of for) as just..(then) so such..that so-that that	because if..(then) when until as soon as so that so that when	5 15 50 2 5 12 1 5
Sentential Relatives	जिससे जो जिसके कारण	with-which which which-of reason	as a result of which as a result of which because of which	5 1 1
Subordinator	पर (-कर -के करके) समय (हुए) के बाद से के पहले के लिए में के कारण	upon do time happening of later with of before of for in of reason	upon after while while while after due to before in order to while because of	9 111 1 28 3 1 1 4 1 3
Coord. Conj.	लेकिन पर परन्तु और तथा या यों TOP..पर ना केवल..बल्कि	but and or such TOP..but not only..but..also	but and or but not only..but also	51 117 2 2 1
Adverbial	तब बाद में फिर इसीलिए नहीं तो तभी तो सो वही यही नहीं	then later in then this-for not then then-only TOP so that this-only not	then later then as a result of this otherwise that (alone) is why so not only that	2 5 4 7 5 1 10 1
TOTAL				472

Subordinating Conjunction

- Independent lexical items
 - introduce finite adverbial subordinate clauses
- Often come paired
- **Arg1** : The main clause argument
- **Arg2**: The subordinate clause

मैं इस सभी धन को राज्य के बादशाह को दे देता क्योंकि
वही समस्त धरती

I this all wealth ACC kingdom of king DAT give give, because he-EMPH all earth
की सम्पदा का स्वामी है
of wealth of lord is

“I would give all this wealth to the king, because he alone is the lord of this whole world’s wealth.”

Subordinating Conjunction

- Paired connectives...

क्योंकि यह तुम्हारी ज़मीन पर मिला है, इसलिए इस धन पर तुम्हारा अधिकार है
because **this your land on found has,** this-for **this treasure on your right is**
“Because this was found on your land, you have the right to this treasure.”

Subordinating Conjunction

■ ki

- Used as a complimentizer
- Can be used as a temporal subordinator too

वह बाल्टी के गंदे पानी से अपनी
चौकलेट निकालने
he bucket of dirty water from his chocolates taking-
out

ही वाला था कि उसकी मम्मी ने उसे रोक दिया
just doing was that his mother ERG him stop did

“He was just going to take out the chocolates from the dirty water in the bucket when his mother stopped him.”

Subordinators

- Elements introducing non-finite subordinate clauses
- Hindi non-finite subordinate clauses almost always appear with overt marking.
 - Post-positions
 - particles following verbal participle
 - suffixes marking serial verbs

मम्मी के मना करने के कारण रामू थोड़ी थोड़ी चौकलेट बड़े अनंद के साथ
mummy of warning doing of reason Ramu little little chocolate big pleasure of with
खा रहा था.
eat being be

“Because of his mother’s warning, Ramu was eating bits of chocolate with a lot of pleasure.”

Subordinators

- kara, wA_huA, etc
 - Typically treated as a serial verb marker,
 - there are cases that show that there are two separate events...
 - judgments can be very subtle
 - Final decisions on initial annotation and evaluation

•Example (a)

अपनी पत्नी से यह सुना कर लकड़हारा बहुत दुखी
हुआ
self wife from this listen do woodcutter much sad became

“Upon hearing this from his wife, the woodcutter became very sad.”

•Example (b)

देखते ही देखते सब बैल भागते हुए
looking EMPH looking all buffalos running happening shed reach did

“Within seconds all the buffalos came running to the shed.”

गोशाला

Coordinating Conjunctions

- Both inter-sentential and intra-sentential

तभी दरवाज़ा खुला और मालकिन आ गई
then-only door opened and wife come went
“Just then the door opened and the wife came in.”

Discourse Adverbials

■ Independent elements

चड़िया जबान कट जाने और मालकिन के ऐसे व्यवहार से डर गई थी. सो वह
bird tongue cut going and wife of this behavior with fear go had. So she
किसी उड़कर चली गई तरह
some manner flying walk went.

“The bird was scared due to her tongue being cut and because of the wife’s behavior. So she somehow flew away.”

Adding the Discourse layer

- Hindi sentence-level dependency annotation (Begum et al, 2008) exists, but is somewhat underspecified for discourse



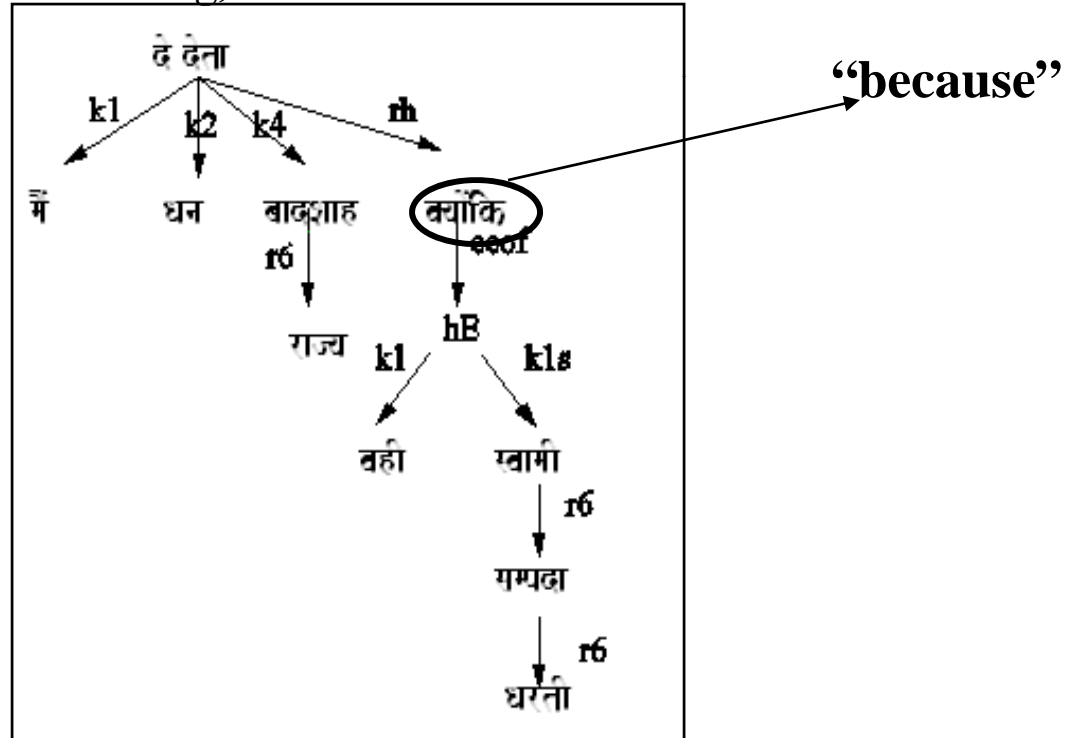
मैं इस सभी धन को राज्य के बादशाह को दे देता

क्योंकि वही समस्त धरती

I this all wealth ACC kingdom of king DAT give give, because he-EMPH all earth

की सम्पदा का स्वामी है
of wealth of lord is

“I would give all this wealth to the king, because he alone is the lord of this whole world’s wealth.”



मैं इस सभी धन को राज्य के बादशाह को दे देता

क्योंकि वही समस्त धरती

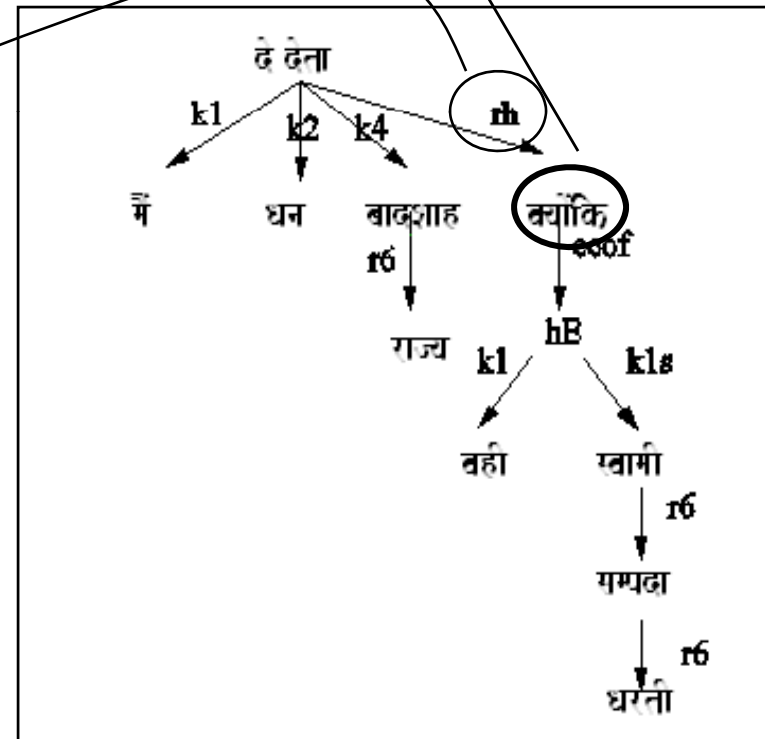
I this all wealth ACC kingdom of king DAT give give, because he-EMPH all earth

की सम्पदा का स्वामी है

of wealth of lord is

“I would give all this wealth to the king, because he alone is the lord of this whole world’s wealth.”

- Dependency level gives the complete analysis
 - Causal relation



मैं इस सभी धन को राज्य के बादशाह को दे देता

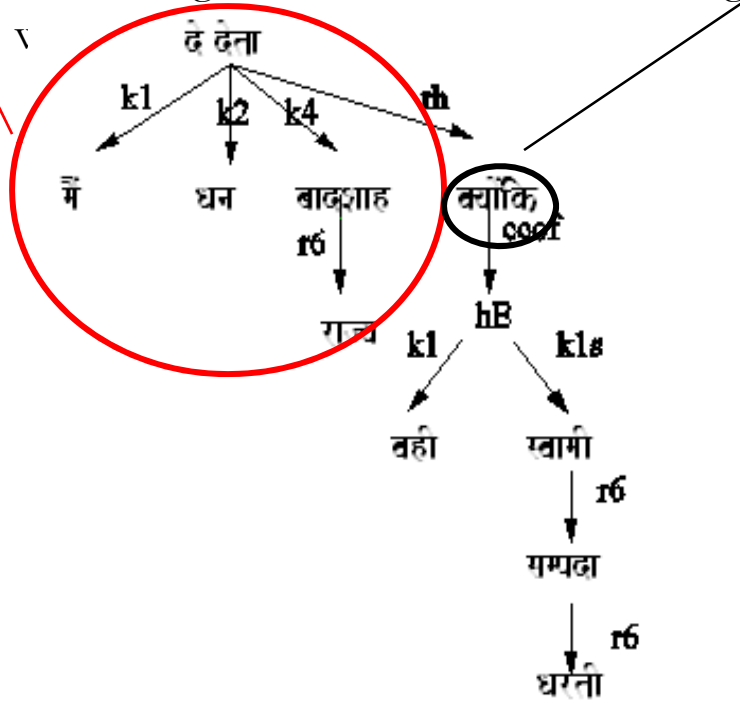
क्योंकि वही समस्त धरती

I this all wealth ACC kingdom of king DAT give give, because he-EMPH all earth

की सम्पदा का स्वामी है

of wealth of lord is

"I would give all this wealth to the king, because he alone is the lord of this whole world's



- Dependency level gives the complete analysis
 - Causal relation

मैं इस सभी धन को राज्य के बादशाह को दे देता

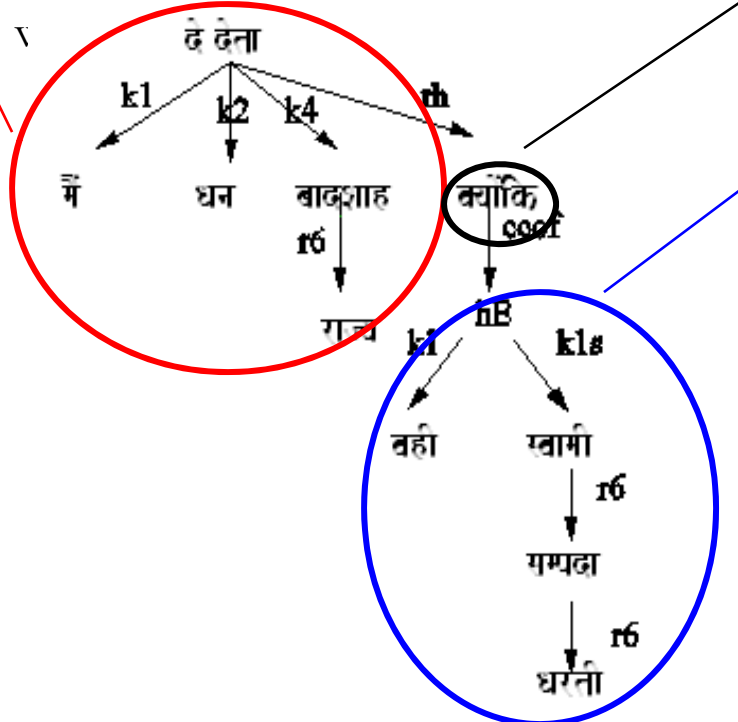
क्योंकि वही समस्त धरती

I this all wealth ACC kingdom of king DAT give give, because he-EMPH all earth

की सम्पदा का स्वामी है

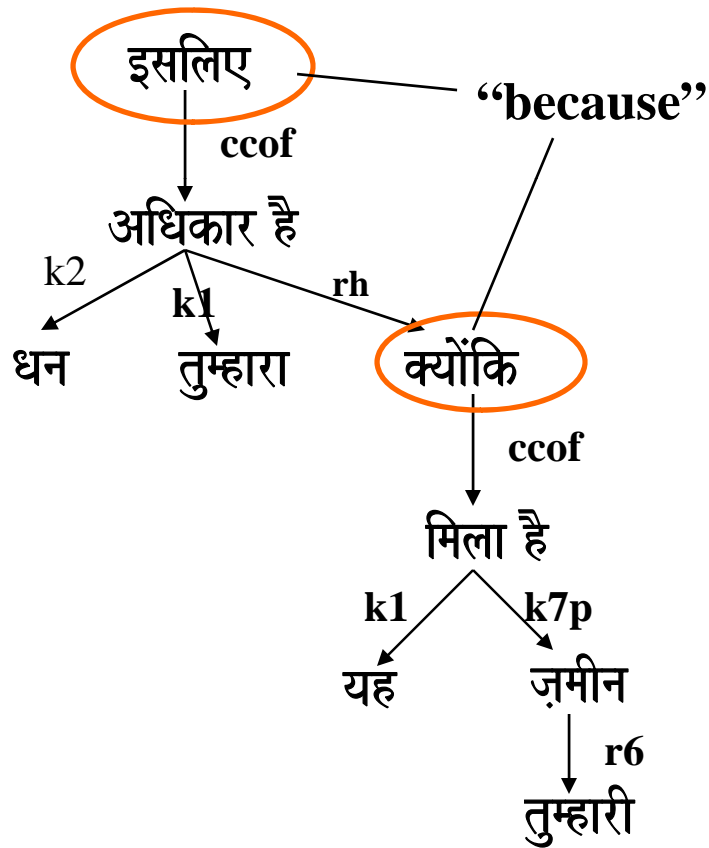
of wealth of lord is

"I would give all this wealth to the king, because he alone is the lord of this whole world's



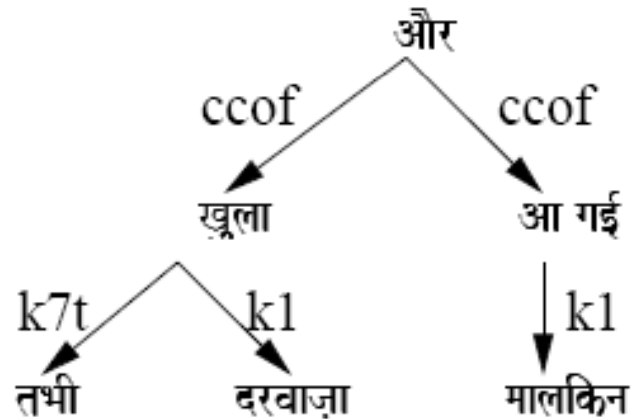
- Dependency level gives the complete analysis
 - Causal relation

क्योंकि यह तुम्हारी ज़मीन पर मिला है, इसलिए इस धन पर तुम्हारा अधिकार है
because this your land on found has, this-for this treasure on your right is
 “Because this was found on your land, you have the right to this treasure.”



- Discourse level annotation
 - Correct argument structure
 - Correct Semantics

तभी दरवाज़ा खुला और मालकिन आ गई ।
then-only door opened and wife come went
“Just then the door opened and the wife came in.”



- Semantics of the Conjunctions under-specified
 - ccof (conjunct of relation)
- Similar scenario: vmods

Need for the discourse layer

- **Need to add the discourse layer over the dependency**
 - Getting the correct argument structure
 - Lexical expression of discourse relations not shown explicitly
 - Semantics is underspecified
 - Discourse layer necessary.
-

Summary

- Initial attempt in exploring different types of discourse connectives in Hindi
 - Language specific cases
 - Motivated the need for the discourse layer
 - Discourse level either underspecified or not present at the dependency layer
-

References

- Rafiya Begum, Samar Husain, Arun Dhvaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of IJCNLP-2008*. Hyderabad, India.
 - The PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS-06-01, IRCS, University of Pennsylvania.
-

Thanks !
