The Hindi/Urdu Treebank: New Frontiers in Hindi and Urdu Natural Language Processing

Dipti Misra Sharma
LTRC, IIIT, Hyderabad, India, dipti@iiit.ac.in

Owen Rambow
CCLS, Columbia, New York City, USA, rambow@ccls.columbia.edu

Ashwini Vaidya
Linguistics, University of Colorado, Boulder, USA, Ashwini.Vaidya@colorado.edu

Dec 8, 2012

COLING 2012

Overview

- Introduction to the nature of syntactic representations. (Rambow, 15 minutes)
- Introduction to the morphology, syntax, and lexical semantics of Hindi and Urdu. (Sharma, 40 minutes)
- The morphological representation for Hindi and Urdu, including encoding issues, tokenization, part-of-speech tags, and morphological representation. (Sharma and Rambow, 20 minutes)
- The dependency representation (DS) for Hindi and Urdu syntax: principles, representation, and examples. (Sharma, 25 minutes)
- The lexical semantic representation (PB) for Hindi and Urdu: principles, representation, and examples. (Vaidya, 25 minutes)
- The phrase structure representation (PS) for Hindi and Urdu syntax: principles, representation, and examples. (Rambow, 25 minutes)
- Sample initial experiments in Hindi and Urdu NLP using the HUTB. (Sharma and Rambow, 15 minutes).

Overview

- Introduction to the nature of syntactic representations. (Rambow, 15 minutes)
- Introduction to the morphology, syntax, and lexical semantics of Hindi and Urdu. (Sharma, 40 minutes)
- The morphological representation for Hindi and Urdu, including encoding issues, tokenization, part-of-speech tags, and morphological representation. (Sharma and Rambow, 20 minutes)
- The dependency representation (DS) for Hindi and Urdu syntax: principles, representation, and examples. (Sharma, 25 minutes)
- The lexical semantic representation (PB) for Hindi and Urdu: principles, representation, and examples. (Vaidya, 25 minutes)
- The phrase structure representation (PS) for Hindi and Urdu syntax: principles, representation, and examples. (Rambow, 25 minutes)
- Sample initial experiments in Hindi and Urdu NLP using the HUTB. (Sharma and Rambow, 15 minutes).

The Hindi Treebank

- 3 Representations
  - DS: Dependency Structure
  - PB: PropBank (lexical predicate-argument structure)
  - PS: Phrase Structure
- Why have three levels of representation? What does “level of representation” mean, in fact?

What is a Syntactic Representation?

1. Syntactic phenomena ("what"), e.g.:
   - Subject of a verb
   - Relative clause
   - Small clause

   Linguists tend to agree on what phenomena exist
2. Mathematical representation type ("basic how"), e.g.:
   - Phrase structure tree
   - Dependency tree
   - Or something more complicated: graph, LFG, TAG, …
3. Formal syntactic description ("detailed how"):
   a. Mapping from phenomena to representations (in particular type)
   b. Chosen representation for a specific phenomenon also called **analysis**
   c. Phenomena extracted in representation are the **interpretation**
   d. Formal description is a **syntactic theory** if it makes predictions

Representation Types: Dependency and Phrase Structure

- Dependency Tree (DS):
  - One label alphabet, words (= words in a sentence)
  - All nodes labeled with words or empty strings
- Phrase Structure Tree (PS):
  - Two disjoint label alphabets, terminals (= words in sentence) and nonterminals
  - All and only interior nodes are labeled with nonterminals
  - Leaves are labeled with terminals or empty strings
- Nothing else is part of the definition!

Example: Small Clauses

- Hindi
  - आतिफ ने सीमा को बेवकूफ समझा
  - Atif ne Seema ko bewakuuf samjhaa
  - Atif Erg Seema Acc stupid consider.Pfv
  - ‘Atif considered Seema stupid.’
- English
  - Atif considered Seema stupid
  - Atif considered her stupid

What is the Phenomenon?

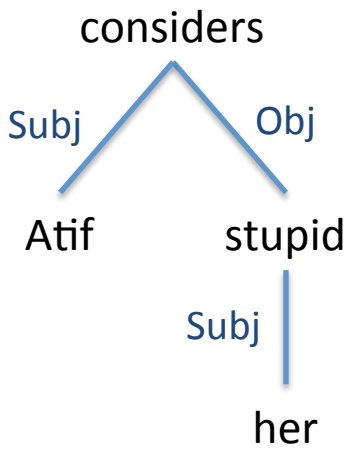
- Syntactically and semantically, *consider* takes a clausal complement
  - Atif considered [$_{\text{clause}}$ that she is stupid]
  - Atif considered [$_{\text{clause}}$ her stupid]
- But two problems:
  - No verb
  - *her* is semantically subject of *stupid* but has accusative case, which is unusual (subjects are usually nominative)
- So:
  - Atif considered [$_{\text{small clause}}$ her stupid]

What is the Representation Type?

- For this example, we will show dependency trees and phrase structure trees

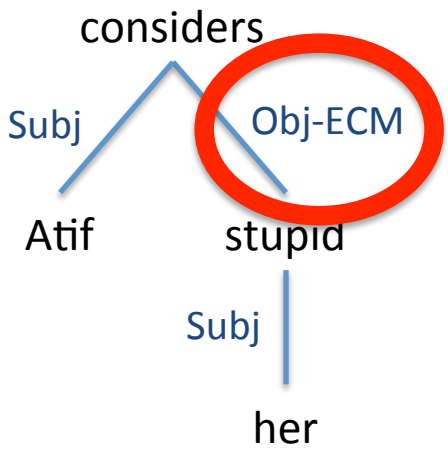
Analysis 1a for Small Clauses: No Accusative Case Marking

- Structure represents *her* as subject but not accusative case marking of *her*



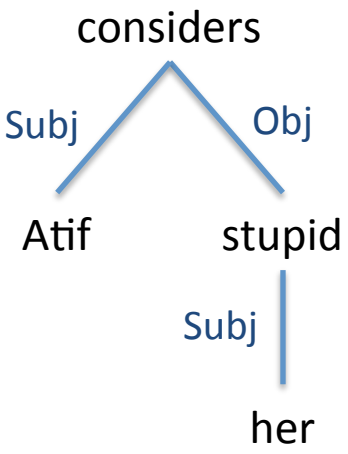
Analysis 1b for Small Clauses: Exceptional Case Marking

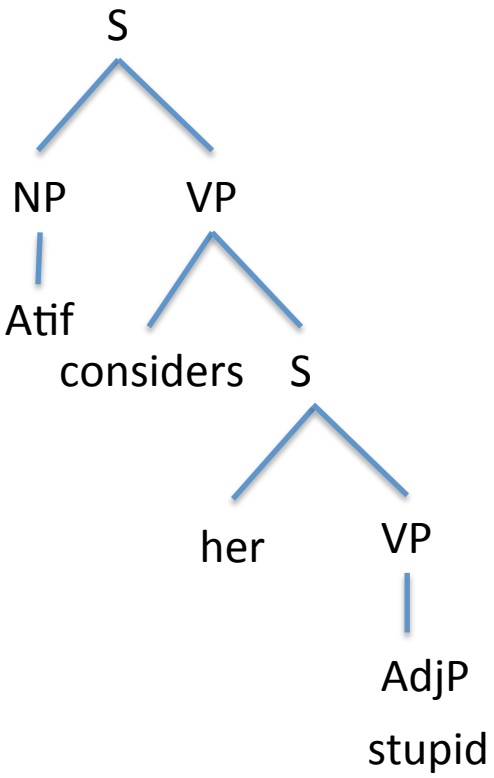
- Structure represents *her* as subject and accusative case marking through node label



Analysis 1a for Small Clauses: No Accusative Case Marking

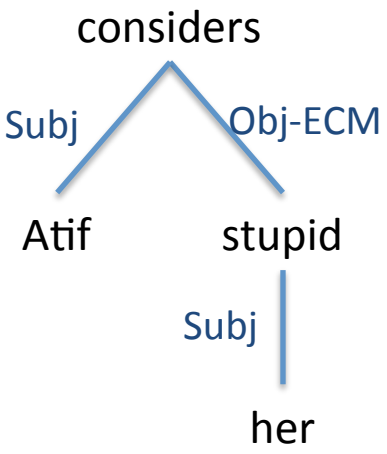
- Structure represents *her* as subject but not accusative case marking of *her*

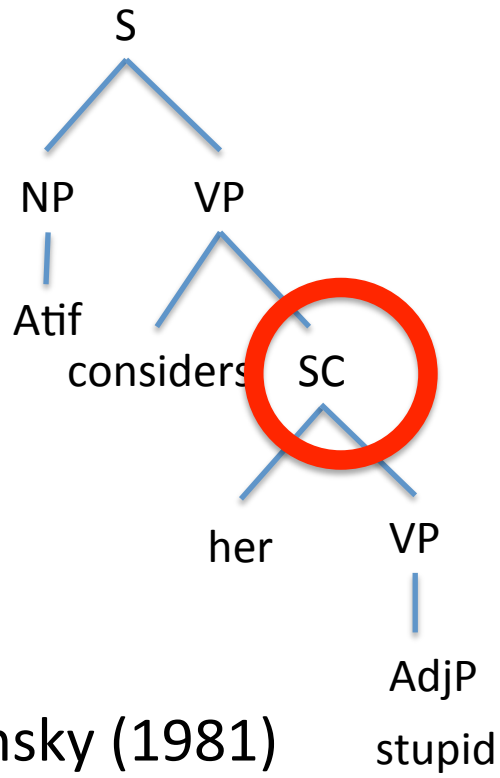




Analysis 1b for Small Clauses: Exceptional Case Marking

- Structure represents *her* as subject but not accusative case marking of *her*





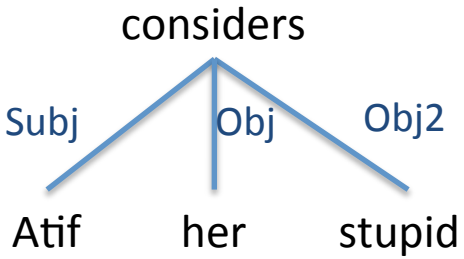
Close to analysis adopted in Chomsky (1981)

Note on DS and PS

- These analyses are intuitively very similar
- Formal notion: “consistency” (Fei Xia, see Bhatt, Rambow & Fei 2011)
  - Intution: very simple and general algorithm can transform consistent DS to PS and *vice versa!*

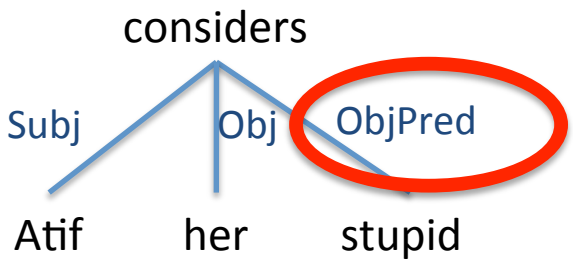
Analysis 2a for Small Clauses: General Monoclausal Analysis

- Structure represents accusative case marking of *her* (as object of matrix verb) but not *her* as semantic subject



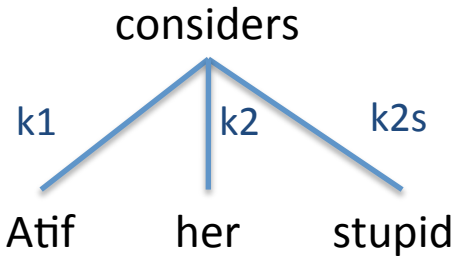
Analysis 2b for Small Clauses: Syntactic Monoclausal Analysis

- Structure represents accusative case marking of *her* (as object of matrix verb) and *her* as semantic subject using node label



Analysis 2b for Small Clauses: Syntactic Monoclausal Analysis

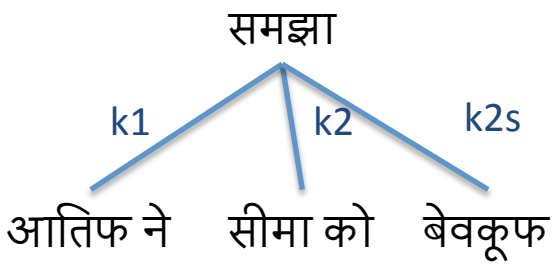
- Structure represents accusative case marking of *her* (as object of matrix verb) and *her* as semantic subject using node label



Neo-Paninian analysis

Analysis 2b for Small Clauses: Syntactic Monoclausal Analysis

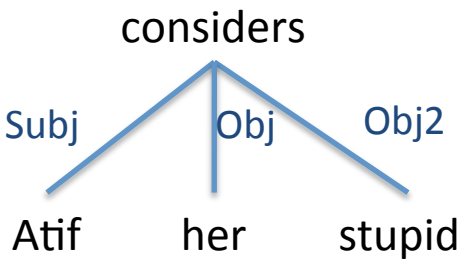
- Structure represents accusative case marking of *her* (as object of matrix verb) and *her* as semantic subject using node label

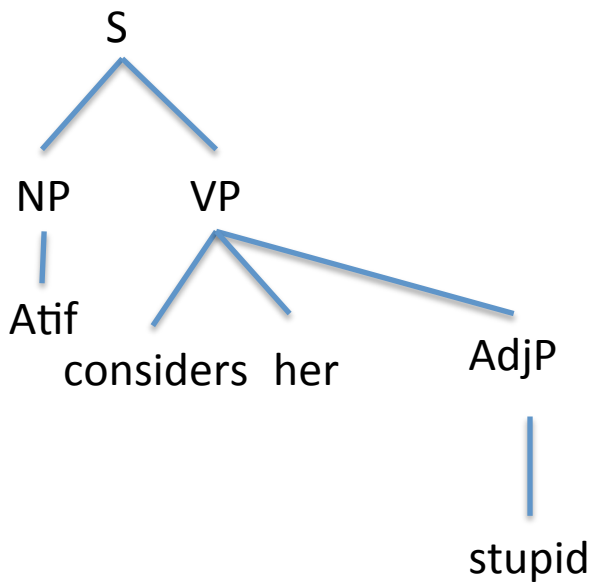


Neo-Paninian analysis from IIIT Hyderabad, Used for DS in Hindi-Urdu Treebank

Analysis 2a for Small Clauses: General Monoclausal Analysis

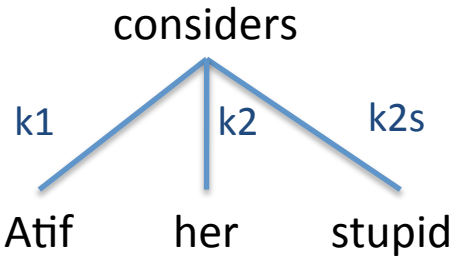
- Structure represents accusative case marking of *her* (as object of matrix verb) but not *her* as semantic subject

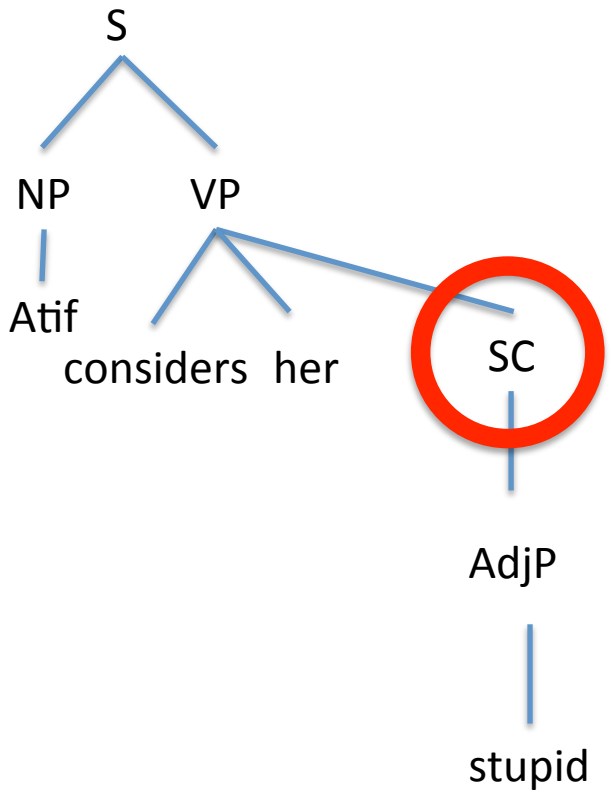




Analysis 2b for Small Clauses: Syntactic Monoclausal Analysis

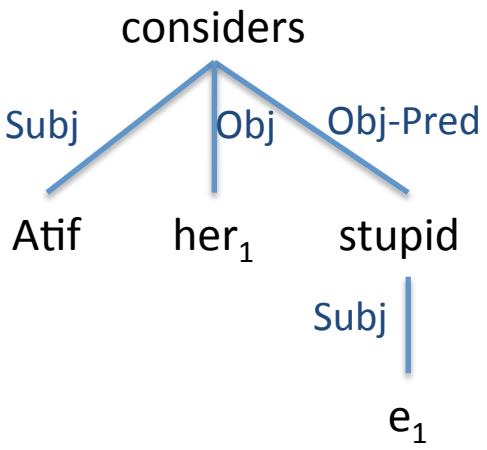
- Structure represents accusative case marking of *her* (as object of matrix verb) and *her* as semantic subject using node label





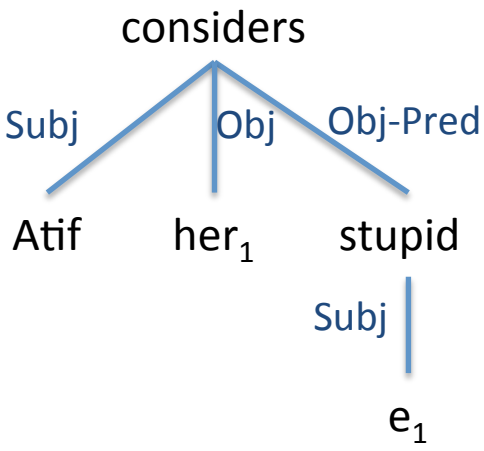
Analysis 3 for Small Clauses: Raising to Object

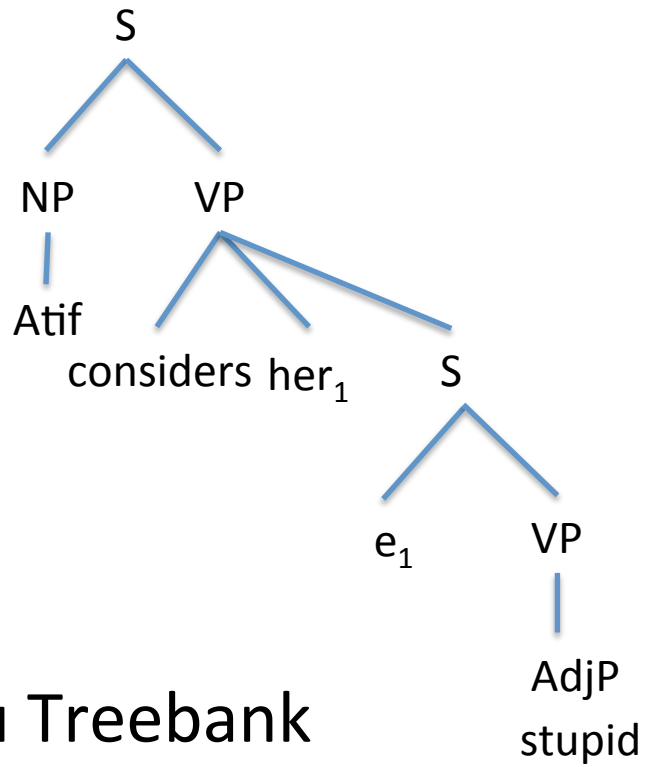
- Structure represents accusative case marking of *her* and *her* as semantic subject but requires empty category



Analysis 3 for Small Clauses: Raising to Object

- Structure represents accusative case marking of *her* and *her* as semantic subject but requires empty category

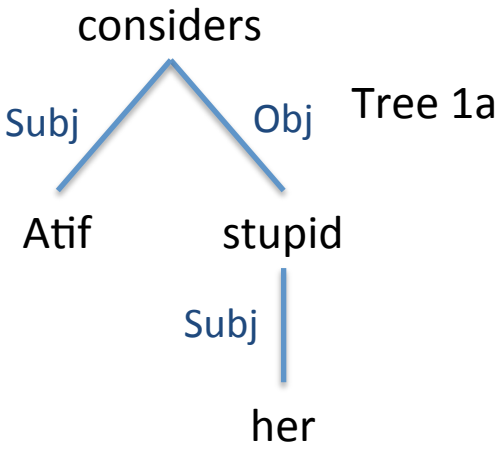




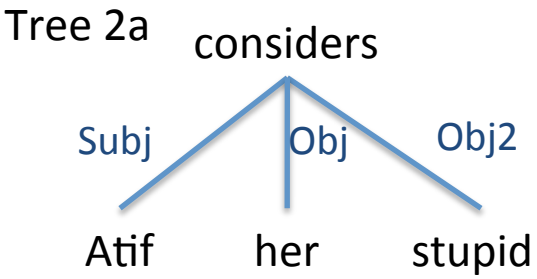
Analysis used for PS in Hindi-Urdu Treebank

Comparison of Representations

- Less Information

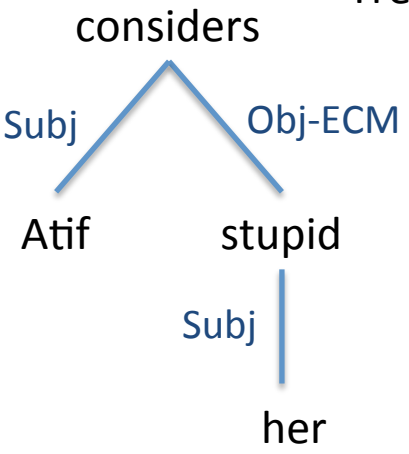


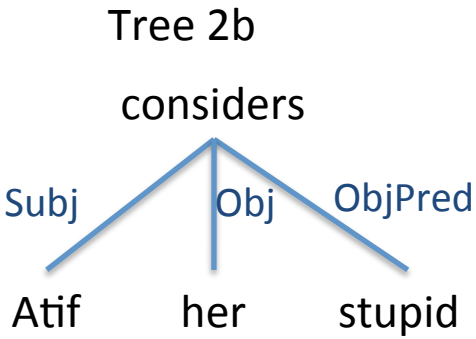
Tree 1a

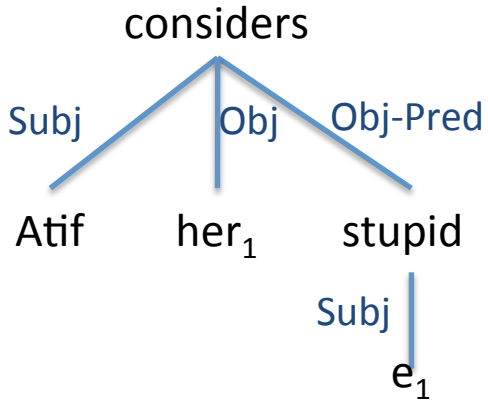


- Same information

Tree 1b







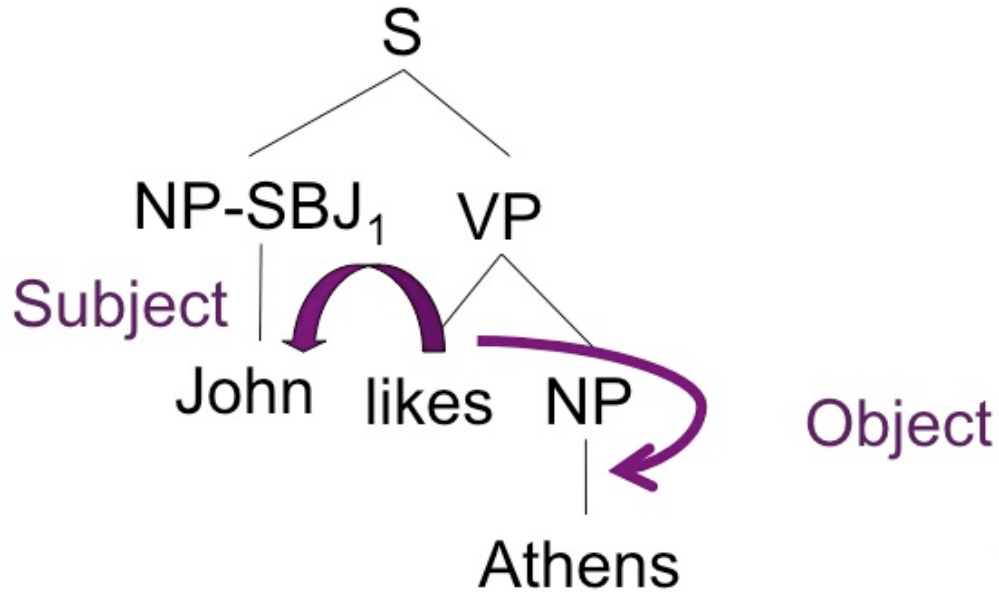
Tree 3

Summary: Syntactic Phenomena, Representation Types, Analyses

- Syntactic phenomena are the empirical data of syntax as part of the science of language
  - Can be very similar across languages
- There can be several possible analyses
  - Some have less information
  - But there can be different analyses that represent the same information differently
- The analyses can be similar in DS and PS
- Lots of choices in treebank design!

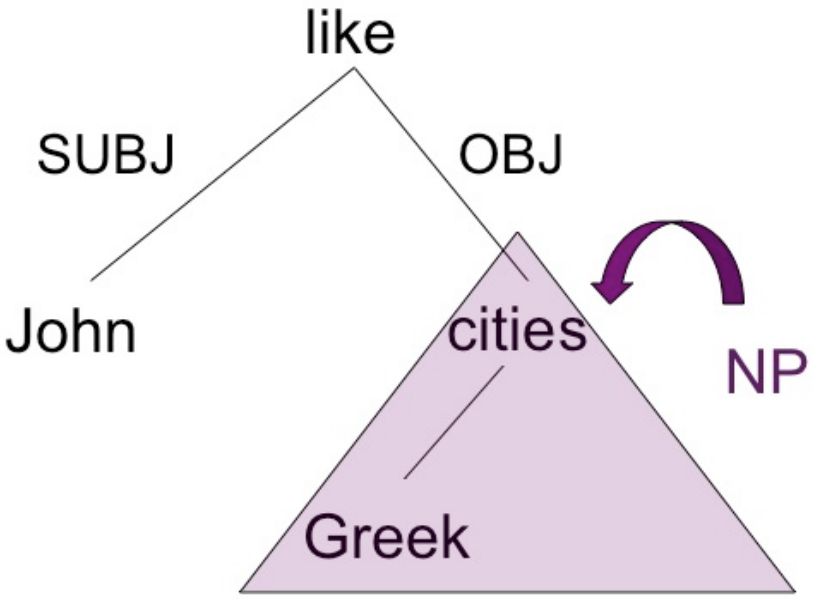
Aren't DS and PS Representations Complementary? NO!

- Syntactic dependency can be encoded in PS, and typically is
- Usual convention: attachment in projection shows type of dependency



Aren't DS and PS Representations Complementary? NO!

- Syntactic constituency is represented in DS
- Usual convention: each node is the word, and the head of the phrase containing it and all descendents



What Does This Mean for NLP?

- Treebanks are not naturally occurring data
- The guidelines are painstakingly produced by linguists and represent a formal description of the language
- Annotators understand a sentence, determine what syntactic phenomena exist, and use the guidelines to choose an analysis for the sentence (a structure)
- Users of the treebank can use the guidelines to interpret the structures and get back the syntactic phenomena present
- These phenomena, and not their representation in the treebank, can be used for NLP in *whatever representation chosen by the researcher*!
- There is already lots of linguistics in our resources, we just need to make use of that linguistic information!

The Hindi Treebank

- DS: dependency, annotated by hand
- PB: annotated by hand on top of DS, adds information about lexical semantics
  - Does not change trees
  - Adds labels to arcs and features to nodes
- PS: phrase structure, derived automatically from DS+PB
  - Contains less information than DS+PB
  - DS and PS contain different information

Comparison of DS, PB, PS (Sample)

		DS	PB	PS
How?	Dependency	✔	✔	
	Phrase Structure			✔
What?	Distinguish unergative/unaccusative		✔	✔
	Distinguish temporal/locative adjuncts	✔	✔	
	Distinguish unaccusative/transitive with empty agent	✔		✔

Overview

- Introduction to the nature of syntactic representations. (Rambow, 15 minutes)
- Introduction to the morphology, syntax, and lexical semantics of Hindi and Urdu. (Sharma, 40 minutes)
- The morphological representation for Hindi and Urdu, including encoding issues, tokenization, part-of-speech tags, and morphological representation. (Sharma and Rambow, 20 minutes)
- The dependency representation (DS) for Hindi and Urdu syntax: principles, representation, and examples. (Sharma, 25 minutes)
- The lexical semantic representation (PB) for Hindi and Urdu: principles, representation, and examples. (Vaidya, 25 minutes)
- The phrase structure representation (PS) for Hindi and Urdu syntax: principles, representation, and examples. (Rambow, 25 minutes)
- Sample initial experiments in Hindi and Urdu NLP using the HUTB. (Sharma and Rambow, 15 minutes).