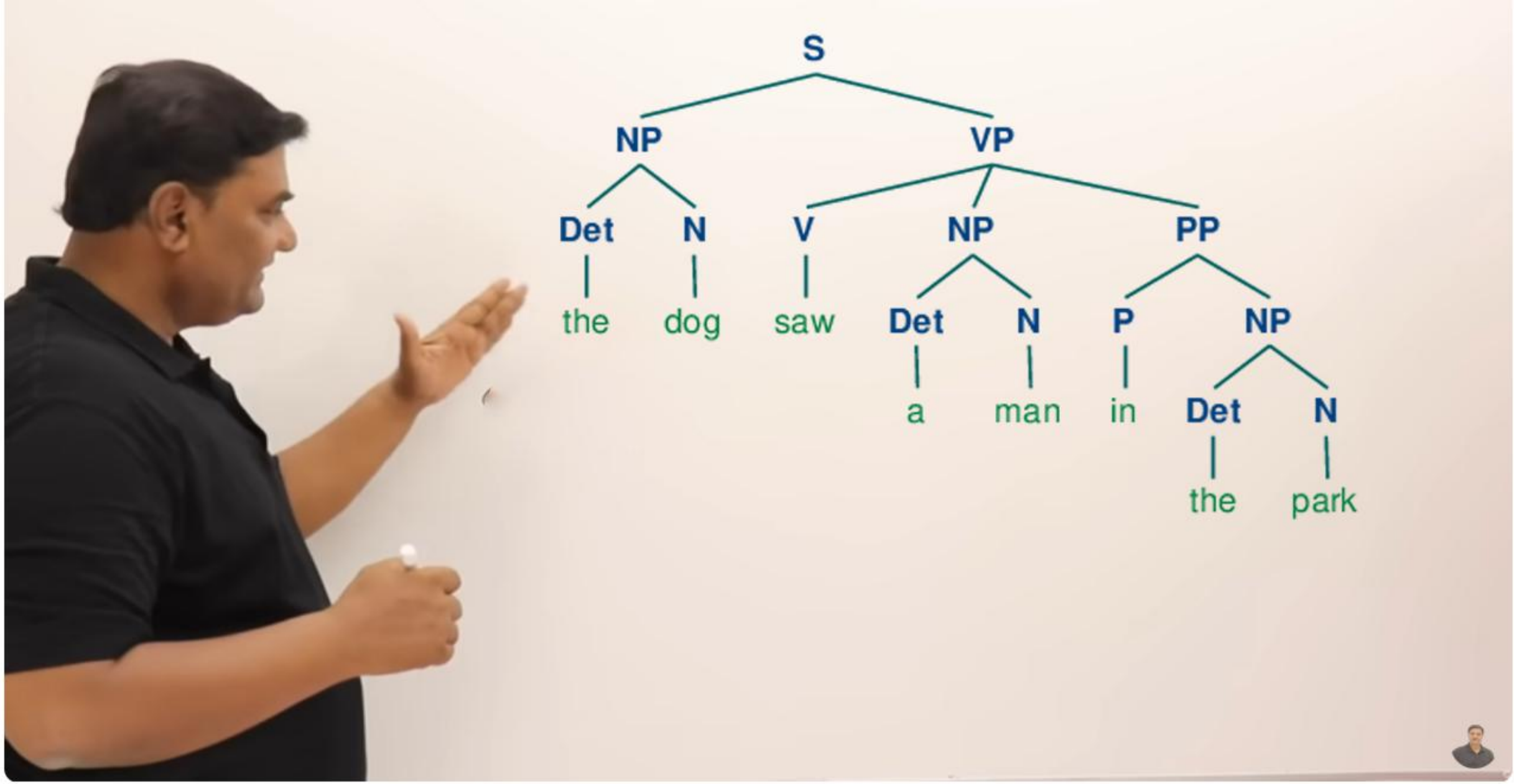
LAB 01: **SYNTACTIC PARSING**

TUTOR: **MINH N.TA**

CLASS FOR THE COURSE OF NATURAL LANGUAGE PROCESSING – IT4772E

SEMESTER 2024.2



Có chàng trai viết lên cây - Phan Mạnh Quỳnh (Mắt Biếc OST)

Abdul Bari
1.12M subscribers

Subscribe

33K | Share | Download | Save | ...

CONTENTS

- Earley's algorithm
- CKY algorithm
- Syntactic parsing using NLTK

EARLEY'S ALGORITHM

EARLEY'S ALGORITHM

```
function Earley-Parse(words,grammar) returns chart
     Enqueue((γ → •S, [0,0]),chart[0])
     for i ← from 0 to Length(words) do
          for each state in chart[i] do
               if Incomplete?(state) and Next-Cat(state) is not POS then
                    Predictor(state)
               elseif Incomplete?(state) and Next-Cat(state) is POS then
                    Scanner(state)
               else
                    Completer(state)
          end
     end
     return(chart)
```

CKY ALGORITHM

CKY ALGORITHM (WITHOUT PROBABILITY)

- Bottom-up parsing: start with the words
- Dynamic programming:
  - save the results in a table/chart
  - re-use these results in finding larger constituents
- Complexity $O(|G|n^3)$ with n: length of string and $|G|$: size of grammar

CKY ALGORITHM (WITHOUT PROBABILITY) – PSEUDOCODE

```
for len = 1 to n:    #number of words in constituent
   for i = 0 to n-len:    #start position
      j = i+len    #end position
      #process unary rules
      foreach A->B where c[i,j] has B
         add A to c[i,j] with a pointer to B
      for k = i+1 to j-1    #mid position
         #process binary rules
         foreach A->B C where c[i,k] has B and c[k,j] has C
            add A to c[i,j] with pointers to B and C
```

SYNTACTIC PARSING USING NLTK

WHAT IS NLTK?

- NLTK (Natural Language Toolkit) is a Python library for processing and analyzing human language.
- Developed in 2001 by Steven Bird and Edward Loper.
- Provides easy-to-use interfaces for over 50 corpora and lexical resources, including WordNet.
- Includes tools for tokenization, parsing, classification, stemming, lemmatization, and more.

KEY FEATURES OF NLTK

- Text Processing: Tokenization, stemming, and lemmatization.
- POS Tagging: Identifies parts of speech in a sentence.
- Named Entity Recognition (NER): Detects names, locations, and other entities.
- Syntax & Semantics: Parsing and grammar processing.
- Text Classification: Sentiment analysis, spam detection, etc.
- Corpus Support: Access to linguistic datasets like Gutenberg, Brown, and Reuters.