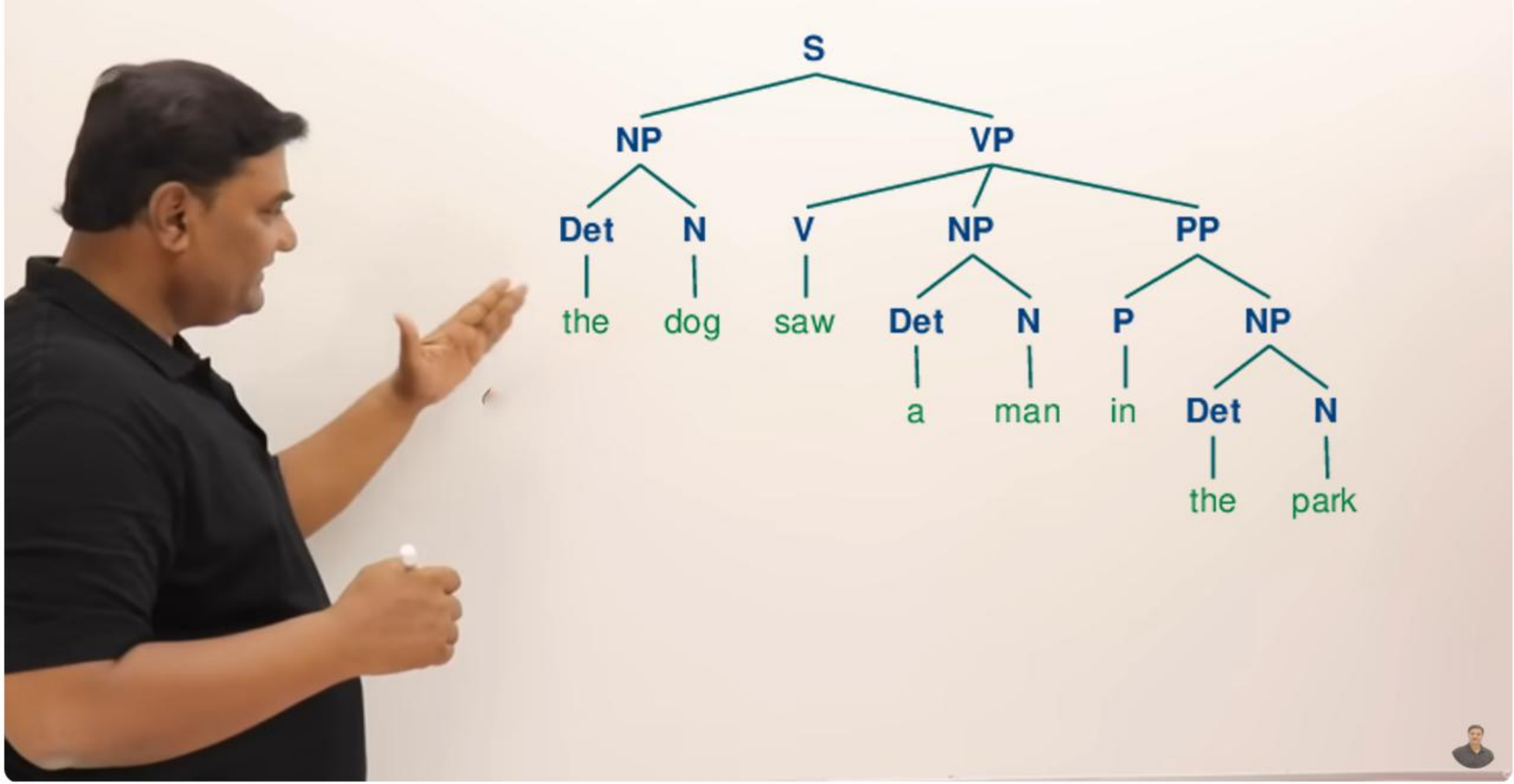
LAB 01: **SYNTACTIC PARSING**

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 | ...

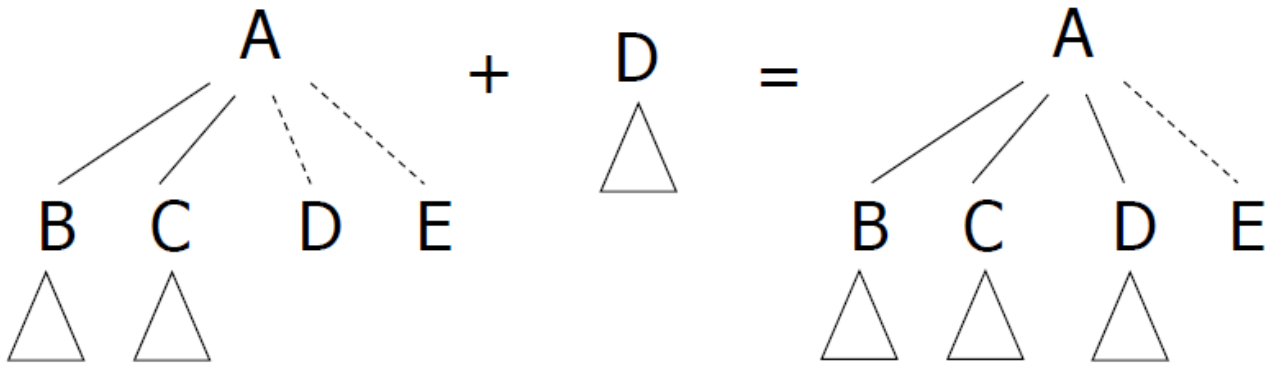
CONTENTS

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

EARLEY'S ALGORITHM

EARLEY'S ALGORITHM

- Finds constituents and partial constituents in input
  - A → B C . D E is partial: only the first half of the A



- Proceeds incrementally, left-to-right

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 i := 1 to n
  - Add to [i-1,i] all categories for the i^{th} word
- for width := 2 to n
  - for start := 0 to n-width
    - Define end := start + width
    - for mid := start+1 to end-1
      - for every constituent X in [start,mid]
      - for every constituent Y in [mid,end]
      - for all ways of combining X and Y (if any)
      - Add the resulting constituent to [start,end] if it's not already there.

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.


THE END
THANK FOR YOUR ATTENTION
makeameme.org