Middle East Technical University Department of Computer Engineering

# CENG 463

## Introduction to Natural Language Processing

Fall ’2012-2013

## Programming Assignment 3

Due date: 30 December 2012, Sunday, 23:55

## 1 Objectives

In this assignment you are expected to build a context free grammar for English with feature structures. You are also expected to write a report on parsing context free grammars including LL and LR parsers and with/without lookahead, recursive descent and shift reduce parsing. You should also explain CKY and Earley Parsers and compare these two.

**Keywords:** *context free grammar, feature structure, unification, parsing*

## 2 Context Free Grammars

Context free grammars are vital in describing sentence structure in natural languages. They are also called phrase structure grammars since they capture the structure of the sentences as combination of phrases. Context free grammars are composed of terminal and nonterminal symbols and rewrite rules to define relations between them.

A formal definition of a CFG is G(V,Σ,R,S) where

V is the set of nonterminal symbols<br>
Σ is the set of terminal symbols<br>
R is the set of rules and defined from V to (VuΣ)*<br>
S is the start symbol

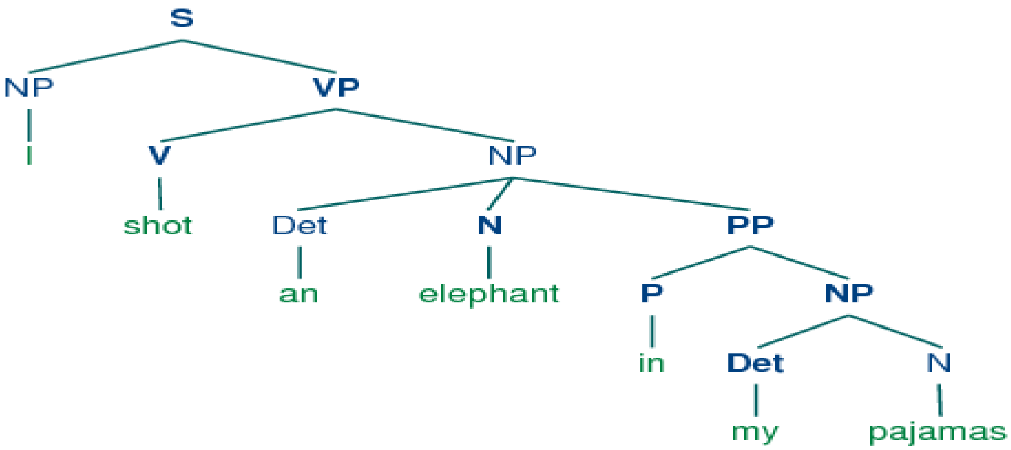
Programming languages are context free languages. Compilers parse text files to match the grammar of the programming language and give syntax error if they do not match.

Sample CFG

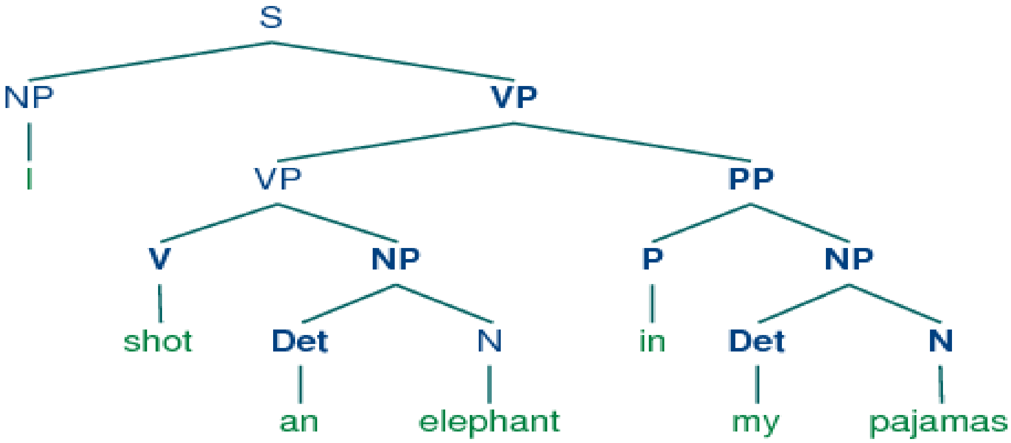
```
S -> NP VP
 PP -> P NP
 NP -> Det N | Det N PP | ’I’
 VP -> V NP | VP PP
 Det -> ’an’ | ’my’
 N -> ’elephant’ | ’pajamas’
 V -> ’shot’
 P -> ’in’
```

Possible derivations of sentence “I shot an elephant in my pajamas”

```
(S
  (NP I)
  (VP
    (V shot)
    (NP (Det an) (N elephant) (PP (P in) (NP (Det my) (N pajamas))))))
```



```
(S
  (NP I)
  (VP
    (VP (V shot) (NP (Det an) (N elephant)))
    (PP (P in) (NP (Det my) (N pajamas)))))
```



# 3 Feature Structures

Feature structures can be seen as attributes of grammar rules. They are functions from features to their values. They provide a way provide complex information in a rule rather than writing multiple rules to describe complex knowledge. They can be used to implement agreement and subcategorization.

Example grammar with feature structures

```
S -> NP[NUM=?n] VP[NUM=?n]
# NP expansion productions
NP[NUM=?n] -> PropN[NUM=?n]
NP[NUM=?n] -> Det[NUM=?n] N[NUM=?n]
NP[NUM=pl] -> N[NUM=pl]
# VP expansion productions
VP[TENSE=?t, NUM=?n] -> IV[TENSE=?t, NUM=?n]
VP[TENSE=?t, NUM=?n] -> TV[TENSE=?t, NUM=?n] NP
Det[NUM=sg] -> 'this' | 'every'
Det[NUM=pl] -> 'these' | 'all'
Det -> 'the' | 'some' | 'several'
PropN[NUM=sg]-> 'Kim' | 'Jody'
N[NUM=sg] -> 'dog' | 'girl' | 'car' | 'child'
N[NUM=pl] -> 'dogs' | 'girls' | 'cars' | 'children'
IV[TENSE=pres,  NUM=sg] -> 'disappears' | 'walks'
TV[TENSE=pres, NUM=sg] -> 'sees' | 'likes'
IV[TENSE=pres,  NUM=pl] -> 'disappear' | 'walk'
TV[TENSE=pres, NUM=pl] -> 'see' | 'like'
IV[TENSE=past] -> 'disappeared' | 'walked'
TV[TENSE=past] -> 'saw' | 'liked'
```

Parse tree of “Kim likes children”

```
(S[]
  (NP[NUM='sg'] (PropN[NUM='sg'] Kim))
  (VP[NUM='sg', TENSE='pres']
    (TV[NUM='sg', TENSE='pres'] likes)
    (NP[NUM='pl'] (N[NUM='pl'] children))))
```

# 4 Parsing

Parsing is the process of resolving a sentence into its component parts of speech with describing relations between them and their syntactic roles. Parsers use grammars to analyze sentences syntactically. There are several approaches in parsing. Top down methods expands grammar rules to match the sentence at some point, bottom up methods combine words reach to the starting symbol. Combining top-down and bottom-up approaches, using dynamic programming, looking multiple words ahead to choose appropriate rules are several examples of enhancements in parsing algorithms.

# 5 Relative Clauses

Relative clauses are modifiers for nouns and noun phrases. They are presented by either one of the relative pronouns or relative adverbs or zero relative. Relative adjectives are which, that, who, whom, whose and adverbs are when, where and why.

Examples:

- Einstein, who was born in Germany, is famous for his theory of relativity.
- Peace is not merely a distant goal that we seek, but a means by which we arrive at that goal.
- Every generation imagines itself to be more intelligent than the one that went before it, and wiser than the one that comes after it.
- The boy, whose parents both work as teachers at the school, started a fire in the classroom.
- My science teacher is a person whom I like very much.

In this assignment, you will only work with ”who” and ”which”.

# 6 Specifications

1. **Y**ou are expected to implement relative clauses in your grammars.
2. **N**ame your grammar extended.fcfg
3. **Y**ou can start with the incomplete example grammar provided then update the grammar based on your needs. You can add or modify rules and update lexicon as needed.
4. **Y**ou may use functions provided with NLTK so please study chapters 8 and 9 of NLP with Python.
5. **Y**ou should provide feature structures for your relative clause rules.
6. **Y**ou should only write rules for relative clauses with ”who” and ”which”.
7. **W**ho is a pronoun used for people and which is used for the rest such as animals and inanimate objects.
8. **B**uild your grammar such that your grammar generates sentences like “the man who. . . ”, “the house which. . . ” etc.
9. **B**uild your grammar such that your grammar does not generate flawed sentences like “the man which. . . ”, “the house who. . . ” etc.
10. **S**tudy examples for further clarification.

# 7 Examples

## 7.1 Positive Examples

- the dog chased the cat which ate the mouse
- people chase Sue who ate the unicorn which Tom saw
- Jody who chases dogs tries to find Tom who likes all cats
- John has a little sister who plays with unicorns
- the park which John hates has unicorns
- Mary saw Sue who ate sandwich
- David ate pizza which Tom gave
- Jody read book which unicorns left at the park
- the fishes eat cats which the mice chase
- the house which Jody likes is very big
- the men who went to the park see the stars with the telescope which the boy find
- people like children who play with unicorns

## 7.2 Negative Examples

- the dog chased who the cat ate the mouse
- people chase Sue which ate the unicorn who Tom saw
- Jody reads books at the park who unicorns play with children
- John has a little sister with what he plays at the park
- David ate pizza which Tom gave pizza
- the house which Jody eats sandwiches is very big
- the dog whom John likes very much disappeared from the park
- some dogs like all cats where eat mice
- a child which chases cats tries to eat sandwiches
- Sue why likes Kim hates Mary
- Tom chases the dog who chases the cat whom chases the mouse
- John whom disappeared from the park lives in a car
- David which he reads books plays in the park
- Jody left the children in the park who the unicorns play
- Kim went to the car whom Mary likes

# 8 Regulations

1. **Programming Language:** You will use Natural Language Toolkit with Python language.
2. **Late Submission:** Penalty for each day is calculated by (number of days) *10.
3. **Cheating:** We have zero tolerance policy for cheating. In case of cheating, all parts involved (source(s) and receiver(s)) get zero. People involved in cheating will be punished according to the university regulations.
4. **R**emember that students of this course are bounded to code of honor and its violation is subject to severe punishment.
5. **Newsgroup:** You must follow the newsgroup (news.ceng.metu.edu.tr) for discussions and possible updates on a daily basis.
6. **Evaluation:** Your grade will be harmonic mean of your precision and recall in other words F1 score.

   $$grade = \frac{2 * precision * recall}{precision + recall} \tag{1}$$

   Precision is the percentage of parsed sentences that are correct over all parsed sentences.
   Recall is the percentage of parsed sentences that are correct over all correct sentences.

   $$precision; = \frac{true\ positives}{true\ positives + false\ positives} \quad recall; = \frac{true\ positives}{true\ positives + false\ negatives} \tag{2}$$

# 9 Submission

- Submission will be done via COW.
  Create a tar.gz file named `hw3.tar.gz` that contains your grammar file 'extended.fcfg' and a copy of your report in pdf format.

# 10 References

- Speech and Language Processing, Daniel Jurafsky and James H. Martin
- Foundations of Statistical Natural Language Processing, Christopher D. Manning and Hinrich Schütze
- Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Loper