Grammar Formalisms for Natural Language Processing

Yoav Goldberg, Fall 2014

this lecture is based on slides
by Julia Hockenmaier
http://cs.illinois.edu/class/fa07/cs498jh

What we will learn?

"linguistics for CS students"

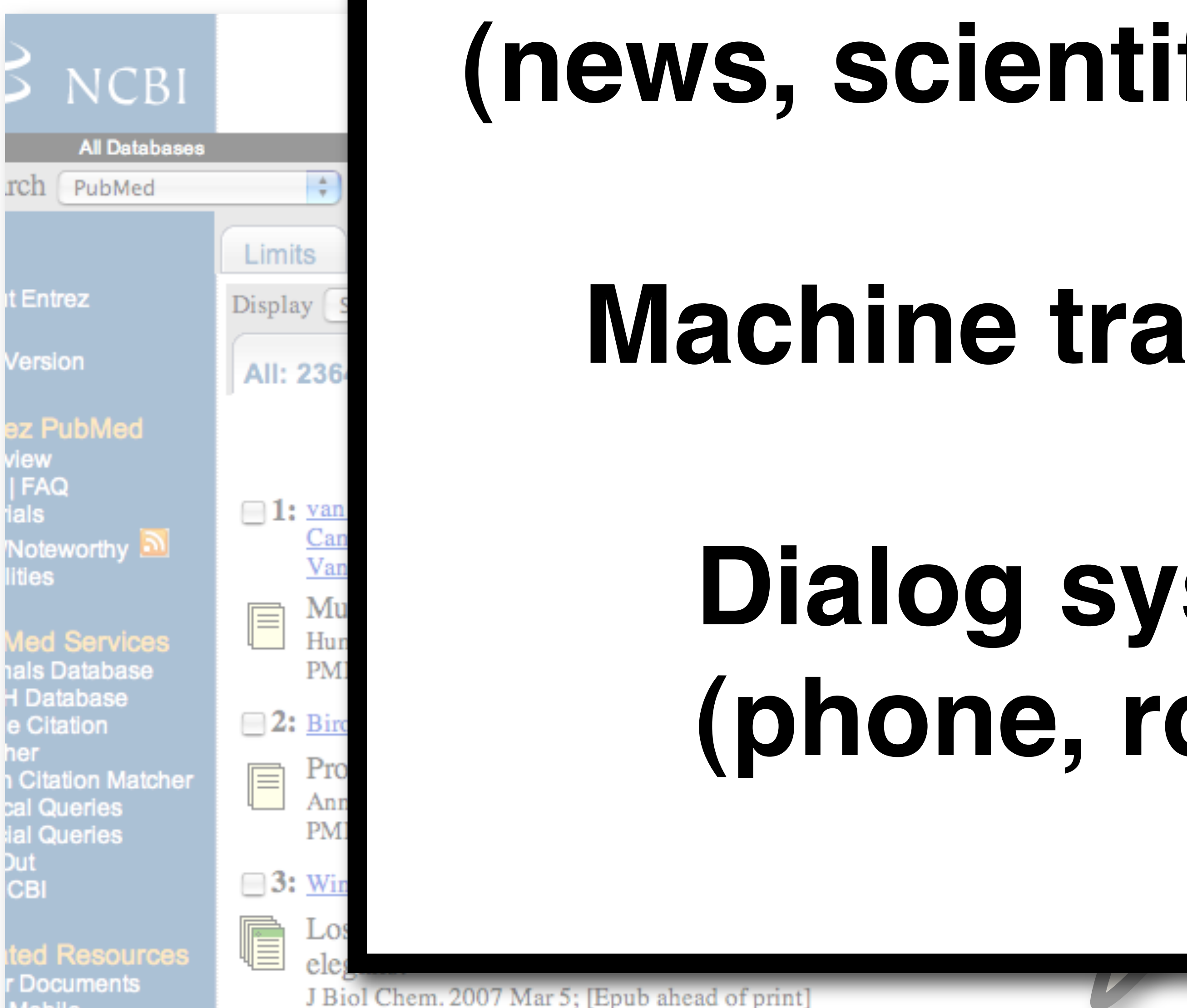
how to represent sentences?
what do we need to represent?

how to use these representations?

how to recover these representations from text?

Why should you take this course?

Natural Language Understanding requires grammars



Information extraction (news, scientific papers)

Machine translation

Dialog systems (phone, robots)

Parsing: a necessary first step

新华社拉萨二月二日电（记者央珍） “八
五”（一九九一至一九九五年）期间，西
藏金融体制改革坚持与全国框架一致、体
制衔接的方针，顺利完成了西藏各级人民
银行的分设工作，实现信贷资金使用从粗
放型经营方式向集约型经营方式转变。

- **What are these symbols?**
  (you need a lexicon)

- **How do they fit together?**
  (you need a grammar)

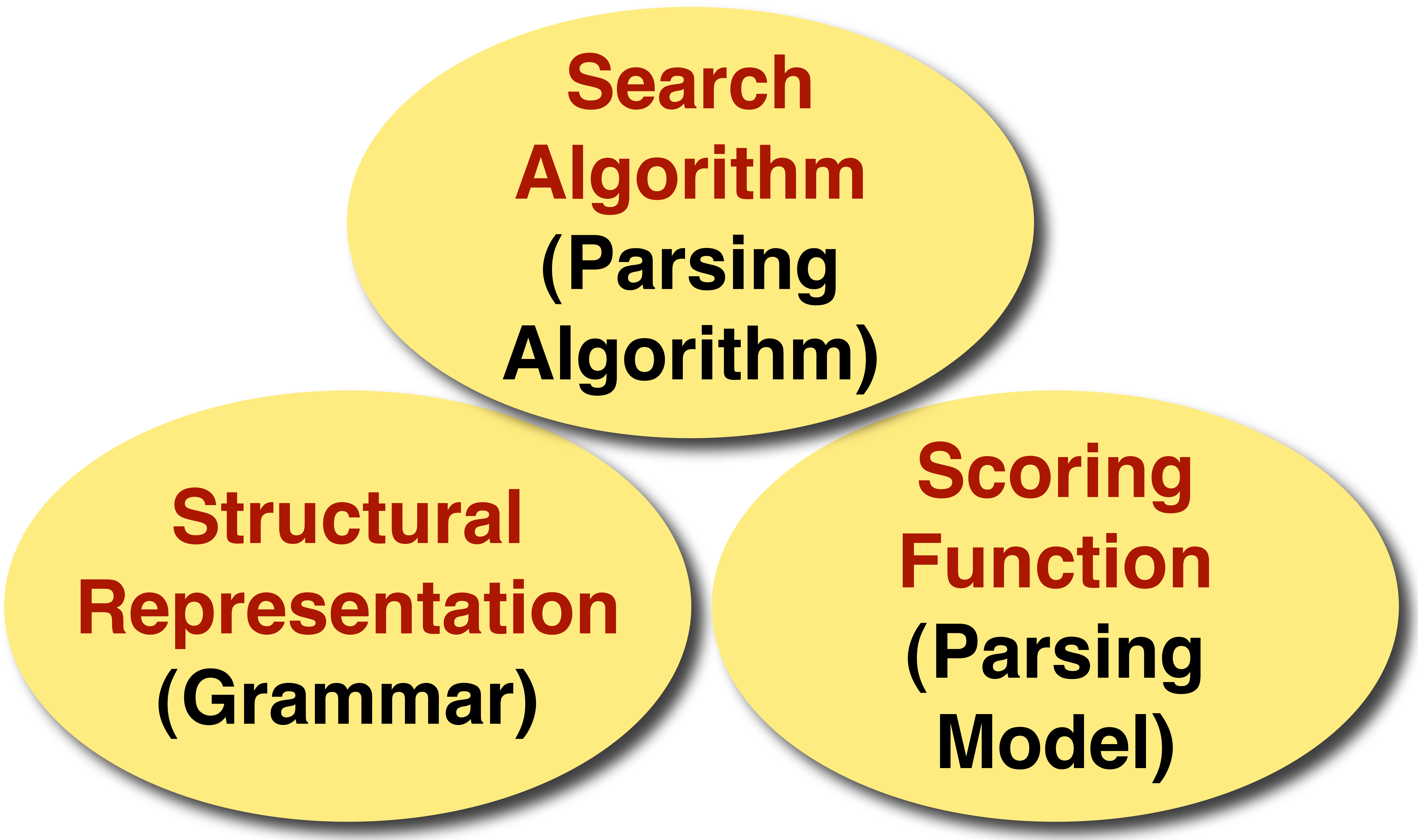
I eat sushi with tuna.

I eat sushi with chopsticks.

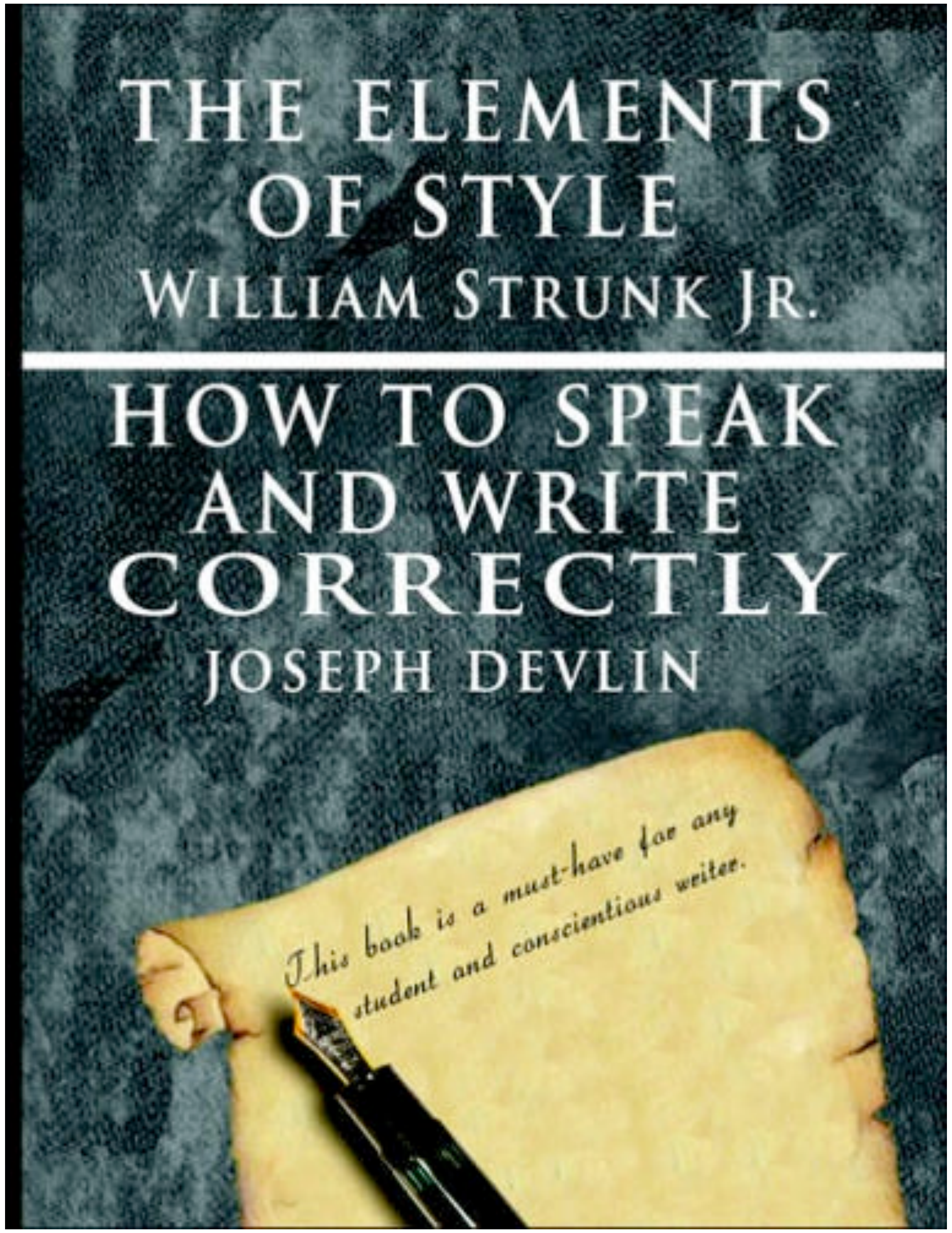
Language is ambiguous.

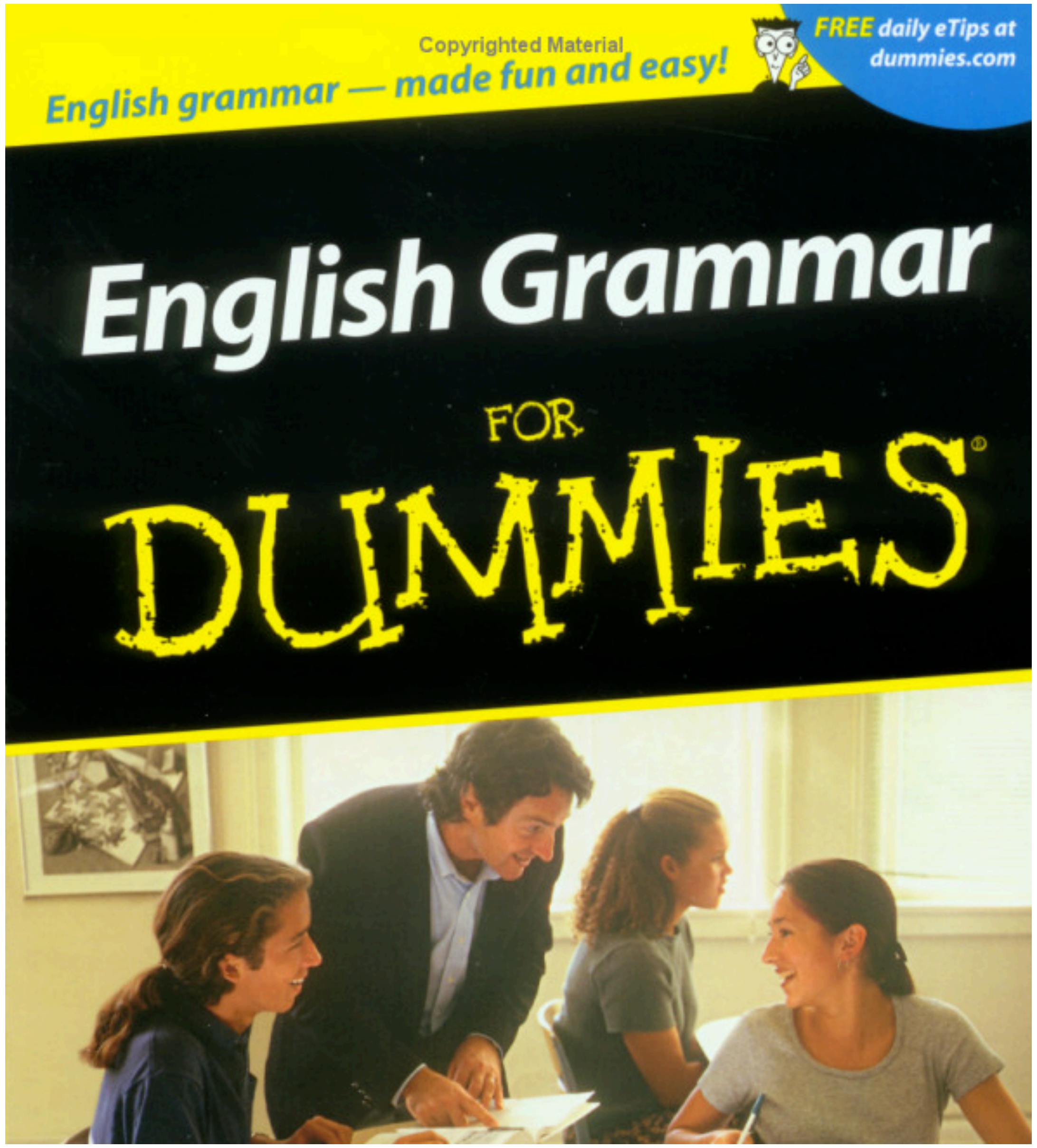
Statistical parsing:
What is the most likely structure?
We need a probability model.

Parsing is a search problem



What is grammar?

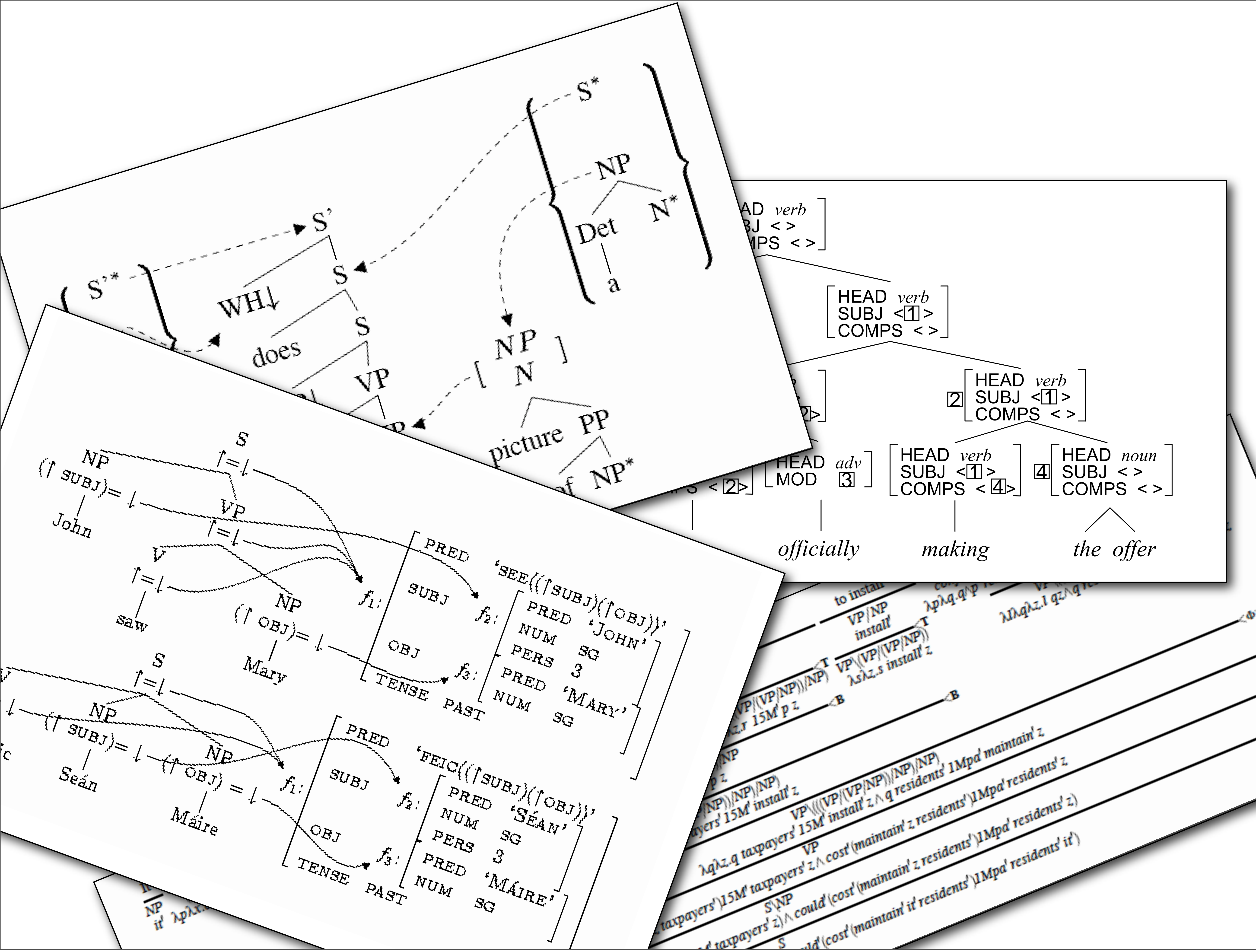




What is grammar?

- **Grammar formalisms**
  (= linguists' programming languages)
  - A precise way to define and describe the structure of sentences.
  - (N.B.: There are many different formalisms out there, which each define their own data structures and operations)

- **Specific grammars**
  (= linguists' programs)
  - Implementations (in a particular formalism) for a particular language (English, Chinese,....)


S*
NP
Det
N*
a
S'
S'*
WH↓
S
does
S
VP
NP
N
picture
PP
NP*
HEAD verb
SUBJ < >
COMPS < >
HEAD verb
SUBJ <1>
COMPS < >
2 HEAD verb
SUBJ <1>
COMPS < >
HEAD adv
MOD 3
HEAD verb
SUBJ <1>
COMPS <4>
4 HEAD noun
SUBJ < >
COMPS < >
officially
making
the offer
S
↑=↓
NP
(↑ SUBJ)=↓
John
VP
↑=↓
V
↑=↓
saw
NP
(↑ OBJ)=↓
Mary
PRED 'SEE⟨(↑SUBJ)(↑OBJ)⟩'
SUBJ
PRED 'JOHN'
NUM SG
PERS 3
OBJ
PRED 'MARY'
NUM SG
TENSE PAST
S
↑=↓
NP
(↑ SUBJ)=↓
Seán
NP
(↑ OBJ) = ↓
Máire
PRED 'FEIC⟨(↑SUBJ)(↑OBJ)⟩'
SUBJ
PRED 'SÉAN'
NUM SG
PERS 3
OBJ
PRED 'MÁIRE'
NUM SG
TENSE PAST

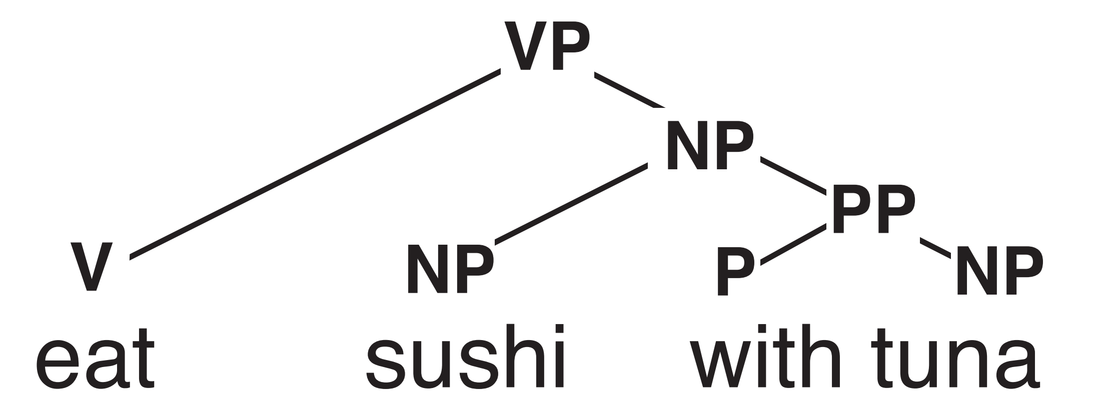
What is the structure of a sentence?

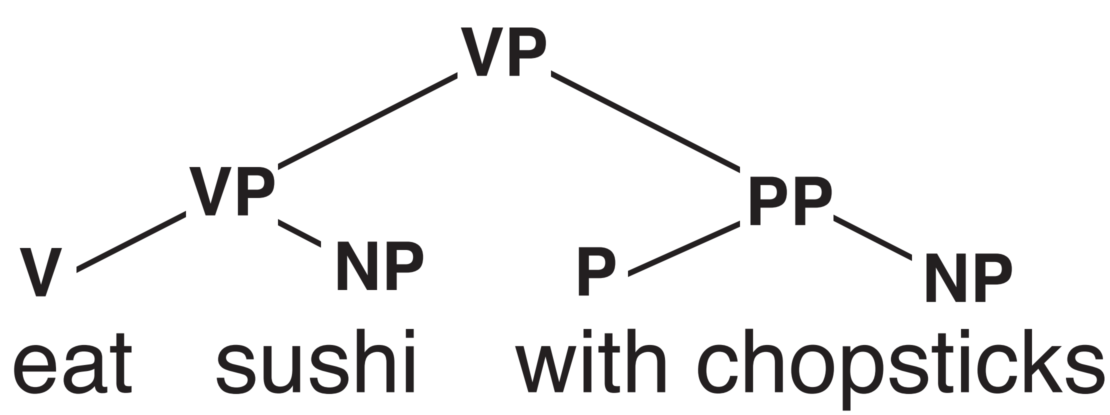
- **Sentence structure is hierarchical:**
  A sentence consists of words (***I, eat, sushi, with, tuna***)
  ..which form phrases: ***"sushi with tuna"***

- **Sentence structure defines dependencies between words or phrases:**

Two ways to represent structure

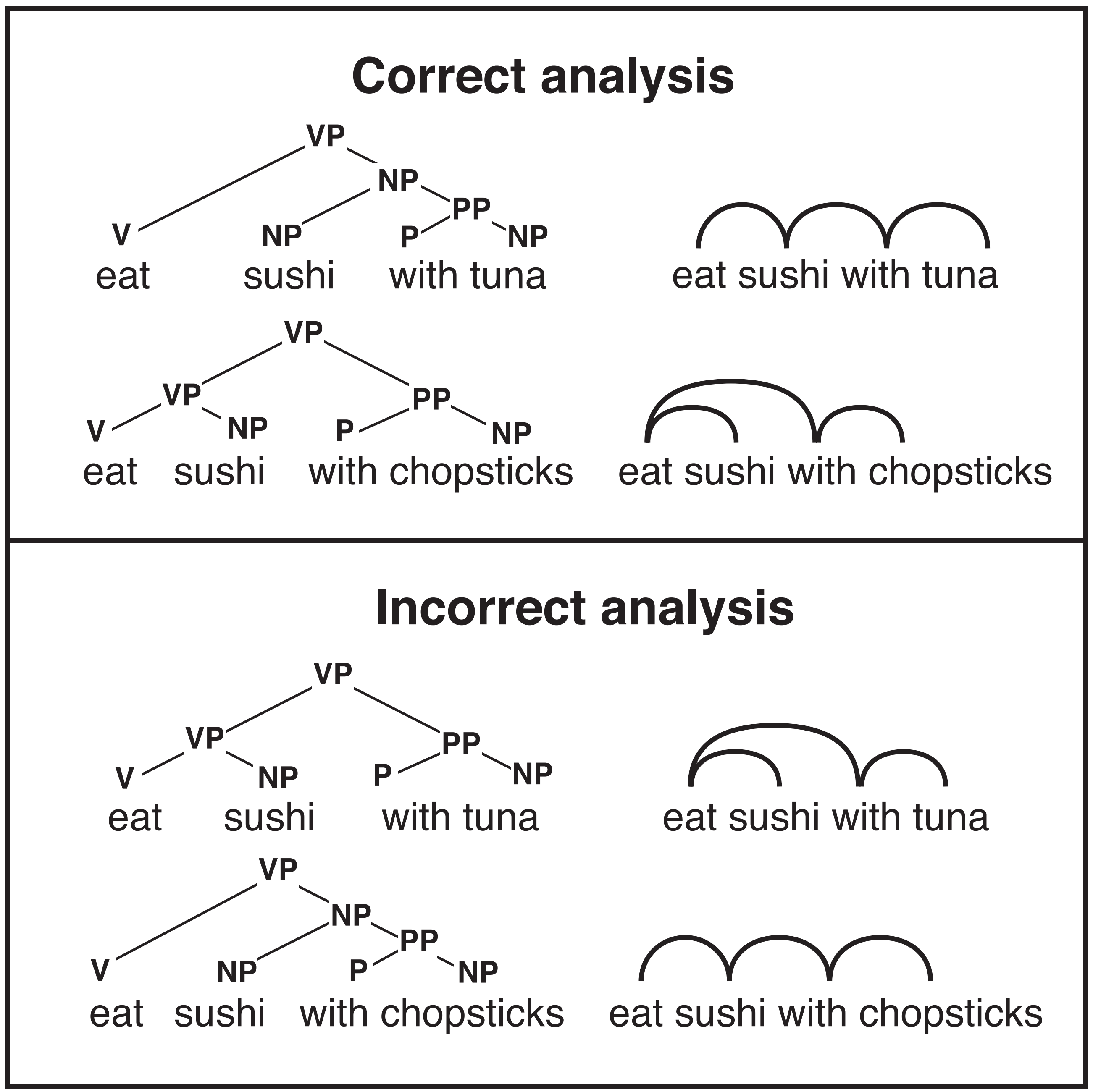
Phrase structure trees





Dependency trees

Structure (Syntax) corresponds to Meaning (Semantics)



What are *expressive* grammar formalisms?

- **They allow richer sets of dependencies.**
  - Context-free grammars: only **nested** dependencies
  - Some languages have **crossing** dependencies.
  - Languages also have additional **non-local** dependencies

Why NLP needs grammars: Machine translation

The output of current systems is often ungrammatical:

Daniel Tse, a spokesman for the Executive Yuan said the referendum demonstrated for democracy and human rights, the President on behalf of the people of two. 3 million people for the national space right, it cannot say on the referendum, the legitimacy of Taiwan's position full.
(BBC Chinese news, translated by Google Chinese to English)

Correct translation requires grammatical knowledge:

"the girl that Mary thinks Jane saw"

- *[das Mädchen], von dem Mary glaubte, dass Jane es gesehen hat.*
- *[la fille] dont Marie croit que Jane l'a vue.*

Why NLP needs grammars: Question Answering

This requires grammatical knowledge...:

John persuaded/promised Mary to leave.

- Who left?

... and inference:

John managed/failed to leave.

- Did John leave?

John and his parents visited Prague. They went to the castle.

- Was John in Prague?
- Has John been to the Czech Republic?
- Has John's dad ever seen a castle?

Research trends in NLP

1980s to mid-1990s: Focus on theory or large, rule-based (‘symbolic’) systems that are difficult to develop, maintain and extend.

Mid-1990s to mid-2000s: We discovered machine learning and statistics! (and nearly forgot about linguistics...oops)
NLP becomes very empirical and data-driven.

Today: Maturation of machine learning techniques and experimental methodology. **We’re beginning to realize that we need (and are able to) use rich linguistic structures after all!**

What will you learn in this course?

Course topics

- **Grammar formalisms**
  - How can you represent the structure of a sentence?
  - How is the same construction represented in different formalisms?
- **Parsing algorithms and models**
  - How can you recover the correct structure of a sentence?
- **Linguistic resources**
  - What data can you use to train a parser?

How does language work?

- ***What sounds are used in human speech?***
  (phonetics)
- ***How do languages use and combine sounds?***
  (phonology)
- ***How do languages form words?***
  (morphology)
- ***How do languages form sentences?***
  (syntax)
- ***How do languages convey meaning in sentences?***
  (semantics)
- ***How do people use language to communicate?***
  (pragmatics)

How does language work?

- ***What sounds are used in human speech?***
  (phonetics)
- ***How do languages use and combine sounds?***
  (phonology)
- ***How do languages form words?***
  (morphology)
- ***How do languages form sentences?***
  (syntax)
- ***How do languages convey meaning in sentences?***
  (semantics)
- ***How do people use language to communicate?***
  (pragmatics)

How does language work?

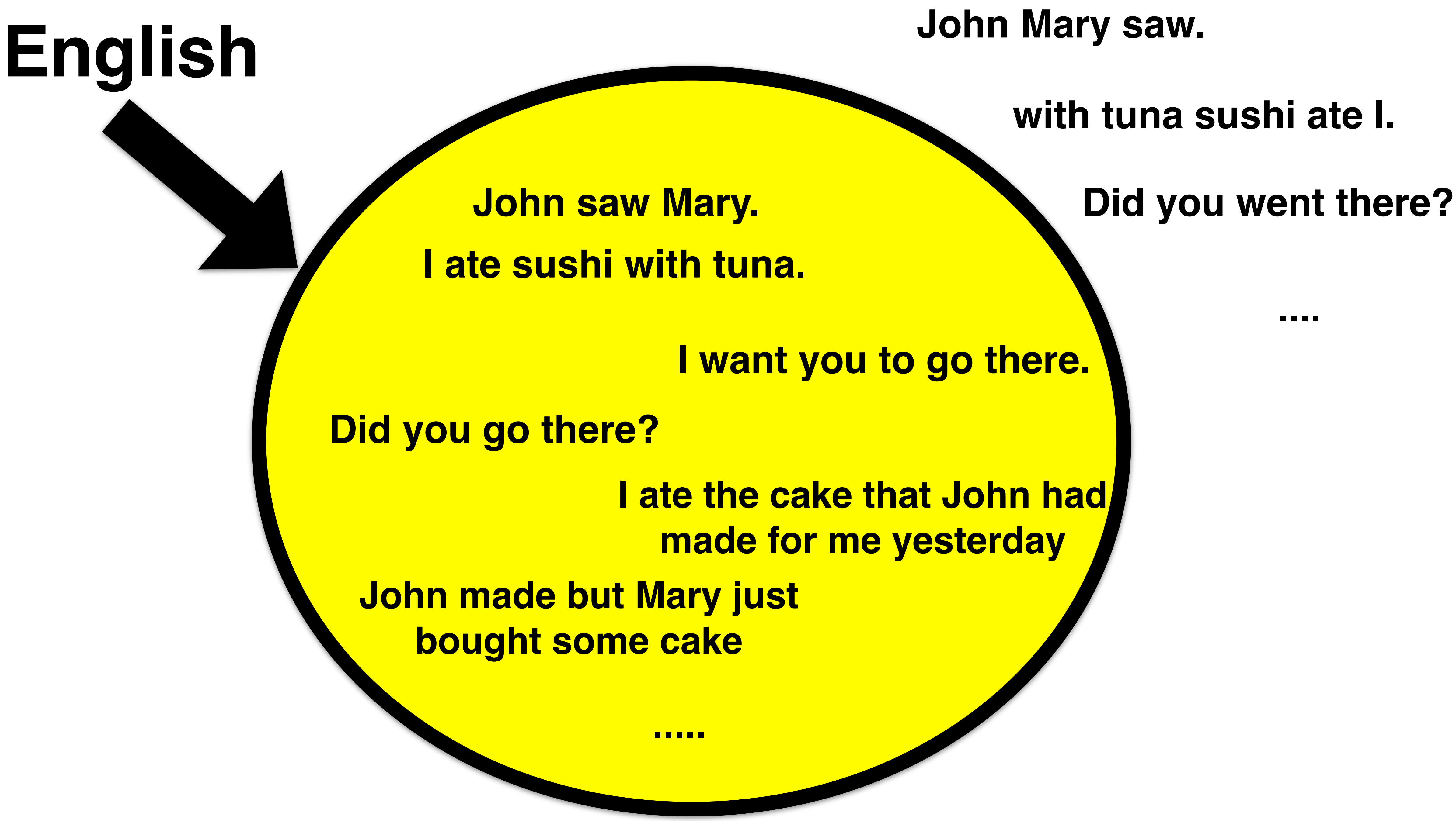
- ***What sounds are used in human speech?* (phonetics)**
- ***How do languages use and combine sounds?* (phonology)**
- ***How do languages form words?* (morphology)**
- ***How do languages form sentences?* (syntax)**
- ***How do languages convey meaning in sentences?* (semantics)**
- ***How do people use language to communicate?* (pragmatics)**

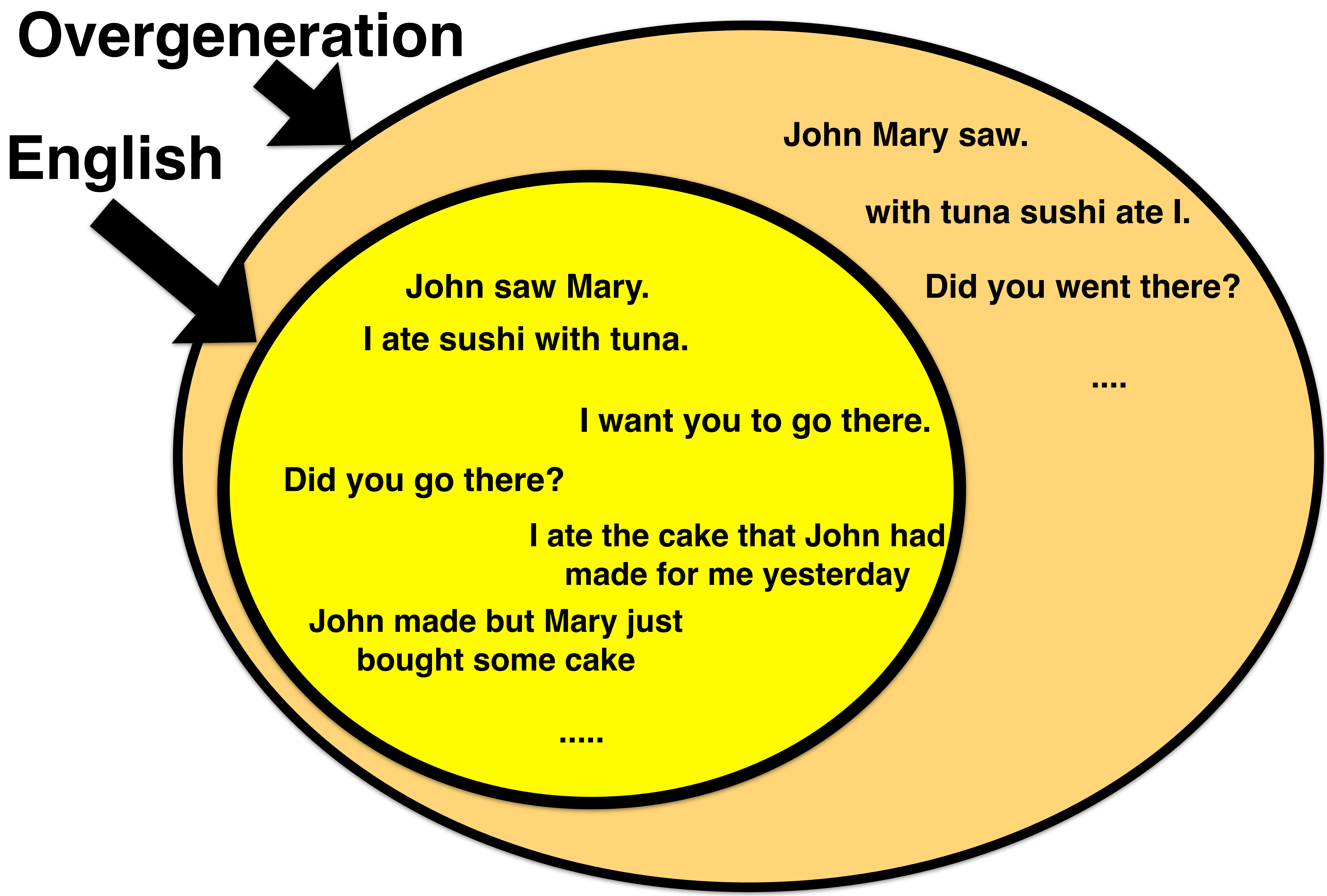
The goal of formal syntax:
Can we define a program that generates all English sentences?

We will call this program "grammar".

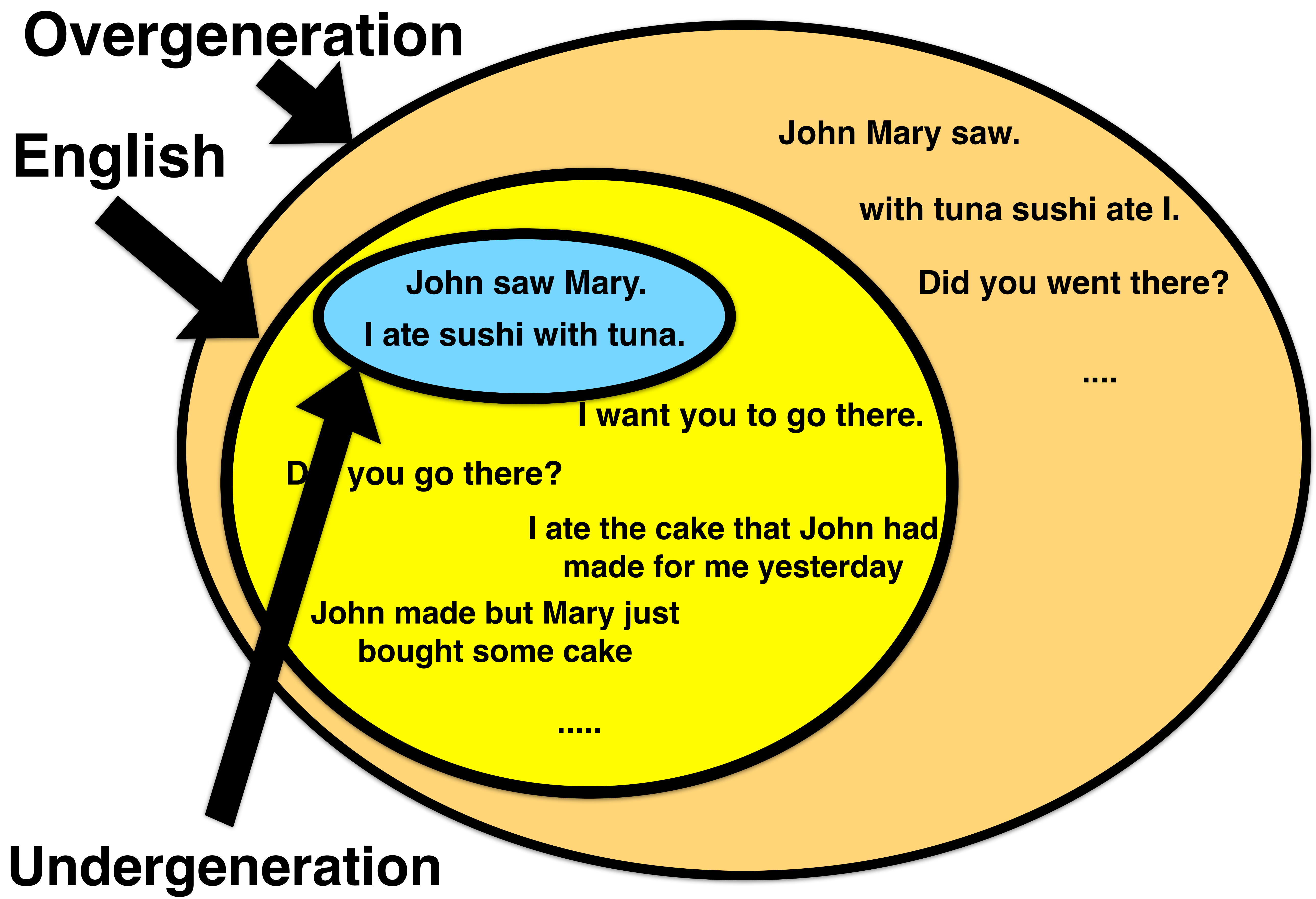
What is the right "programming language" for grammars?

[N.B: linguists demand that the program fit into the mind of a child that learns the language]


English
John saw Mary.
I ate sushi with tuna.
I want you to go there.
Did you go there?
I ate the cake that John had made for me yesterday
John made but Mary just bought some cake
.....
John Mary saw.
with tuna sushi ate I.
Did you went there?
....



Overgeneration
English
John Mary saw.
with tuna sushi ate I.
Did you went there?
....
John saw Mary.
I ate sushi with tuna.
I want you to go there.
Did you go there?
I ate the cake that John had made for me yesterday
John made but Mary just bought some cake
.....


Overgeneration
English
John Mary saw.
with tuna sushi ate I.
Did you went there?
....
John saw Mary.
I ate sushi with tuna.
I want you to go there.
Did you go there?
I ate the cake that John had made for me yesterday
John made but Mary just bought some cake
.....
Undergeneration

Basic word classes (parts of speech)

- **Content words (open-class):**
  - **nouns**: *student, university, knowledge*
  - **verbs**: *write, learn, teach,*
  - **adjectives**: *difficult, boring, hard, ...*
  - **adverbs**: *easily, repeatedly,*

- **Function words (closed-class):**
  - **prepositions**: *in, with, under,*
  - **conjunctions**: *and, or*
  - **determiners**: *a, the, every*

Basic sentence structure

I eat sushi.

Basic sentence structure

I eat sushi.

Noun
(Subject)

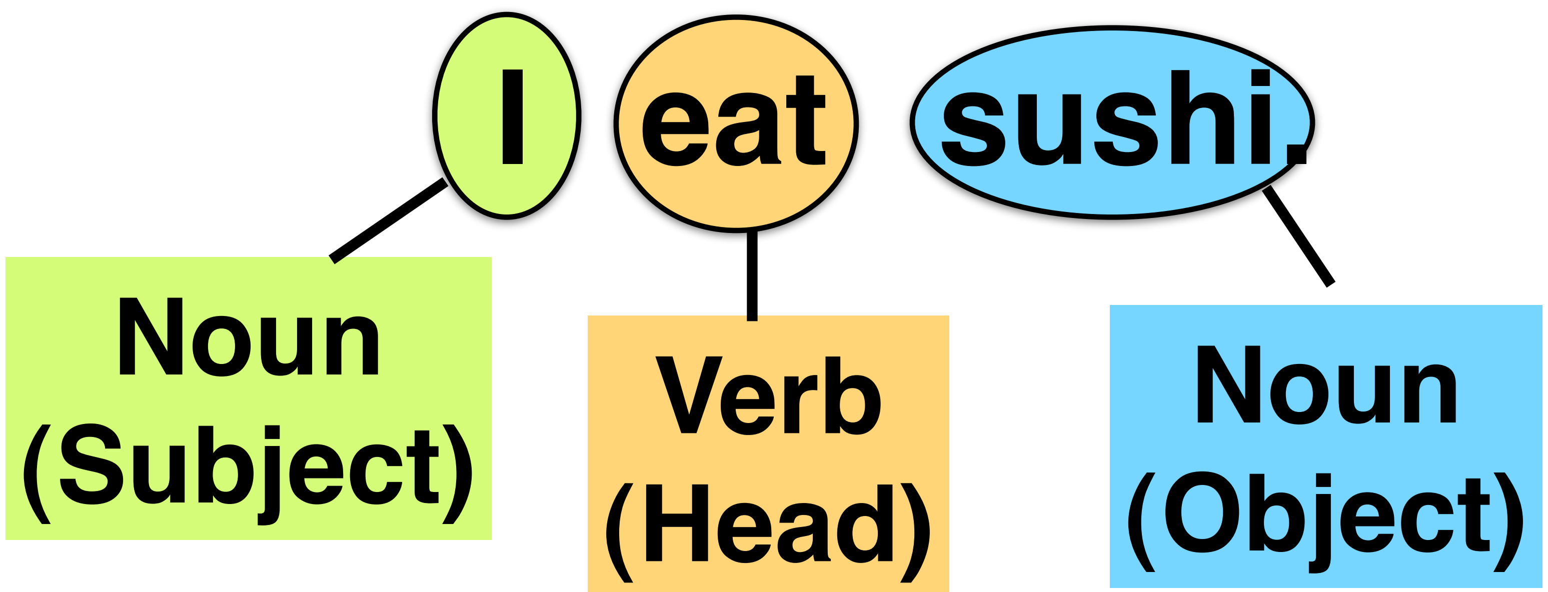
Basic sentence structure

I eat sushi.

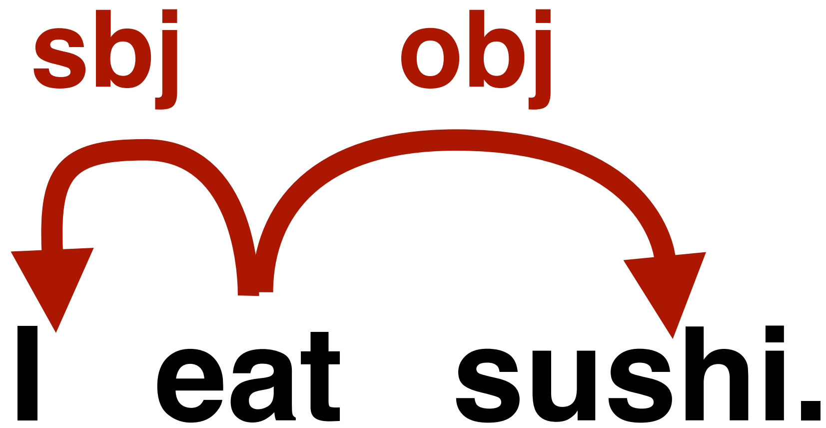
Noun
(Subject)

Noun
(Object)

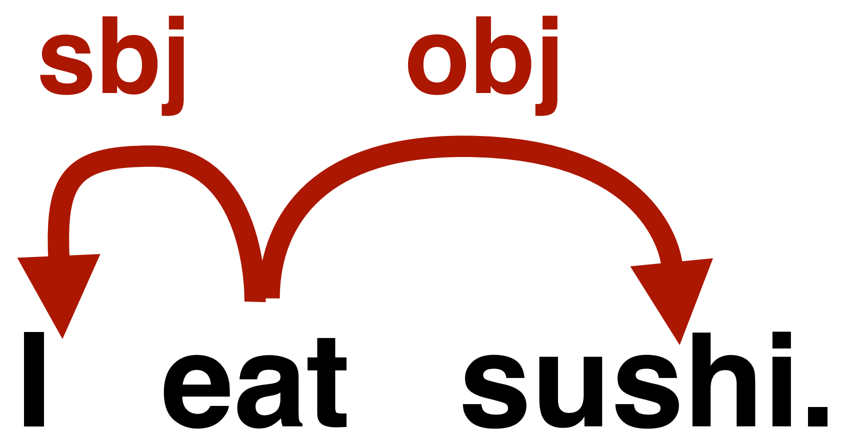
Basic sentence structure

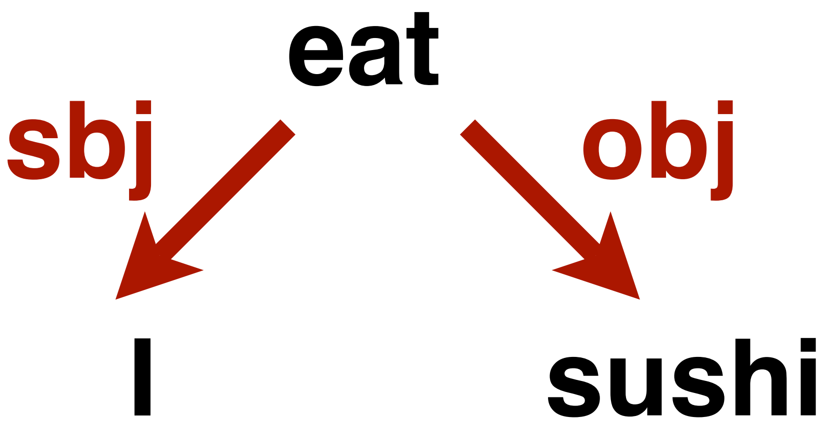


As a dependency tree



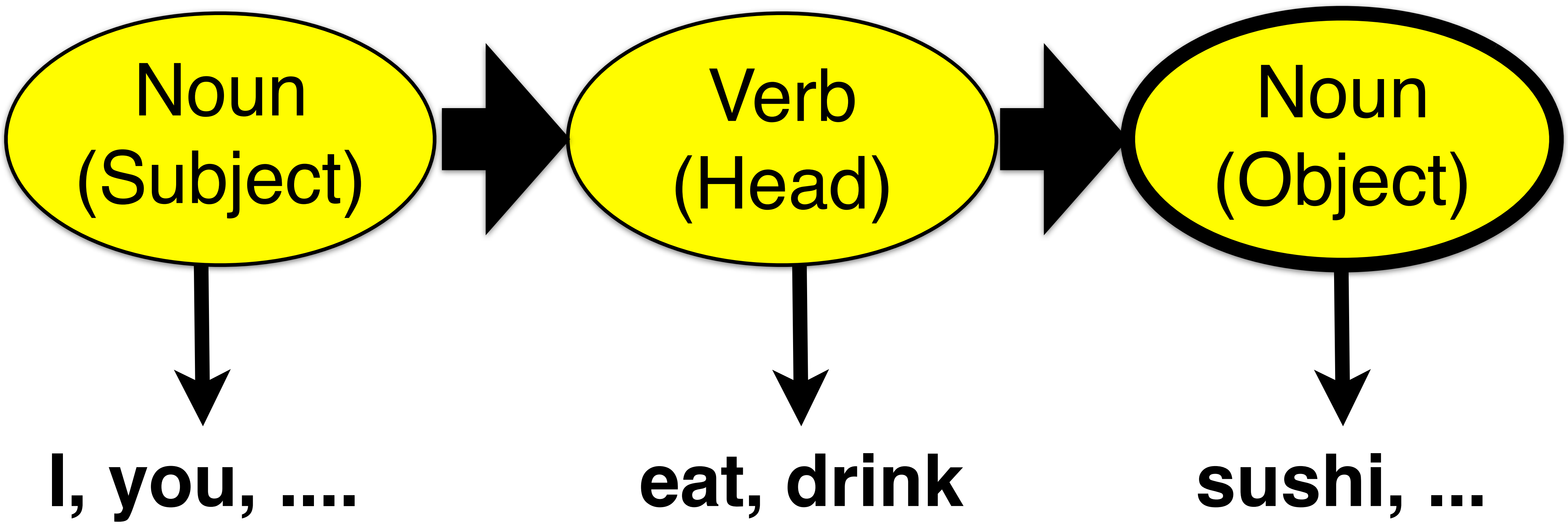
As a dependency tree





A finite-state-automaton (FSA)
(or Markov chain)

A Hidden Markov Model (HMM)



Words take arguments

I eat sushi. ✔
I eat sushi you. ???
I sleep sushi ???
I give sushi ???
I drink sushi ?

Words take arguments

I eat sushi. ✔
I eat sushi you. ???
I sleep sushi ???
I give sushi ???
I drink sushi ?

Subcategorization:
Intransitive verbs (sleep) take only a subject.
Transitive verbs (eat) take also one (direct) object.
Ditransitive verbs (give) take also one (indirect) object.

Selectional preferences:
The object of *eat* should be edible.

A better FSA

Language is recursive

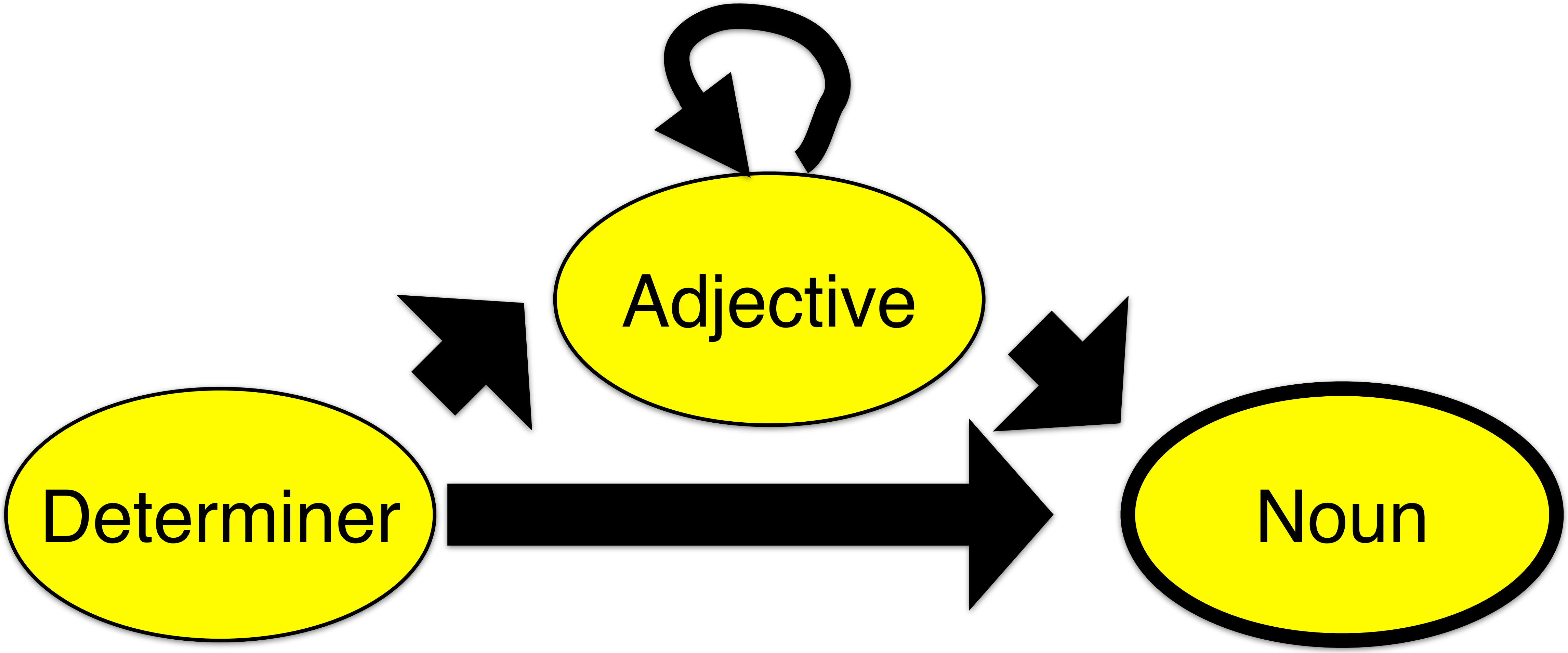
the ball
the big ball
the big, red ball
the big, red, heavy ball
....

Adjectives can **modify** nouns**.**
The **number of modifiers/adjuncts** a word can have is (in theory) **unlimited**.

Can we define a program that generates all English sentences?

The number of sentences is infinite.
But we need our program to be finite.

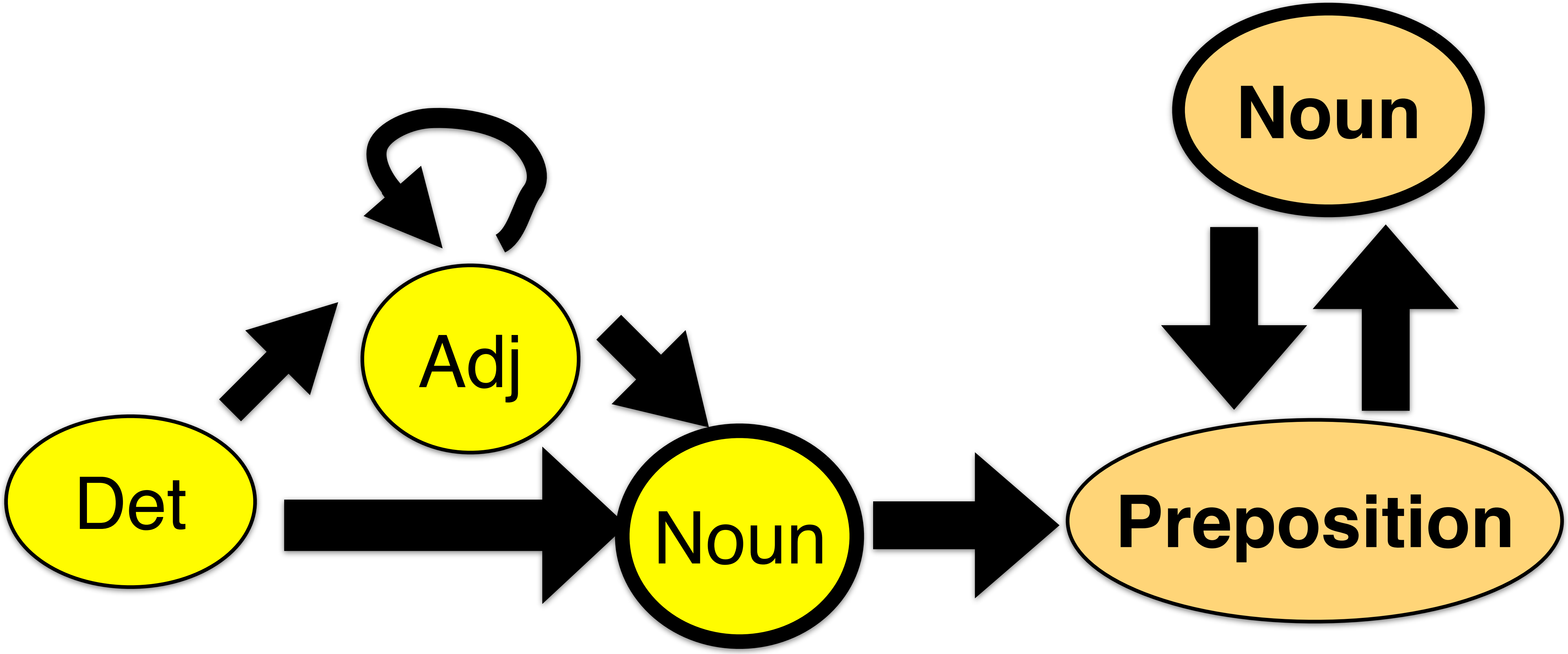
Another FSA



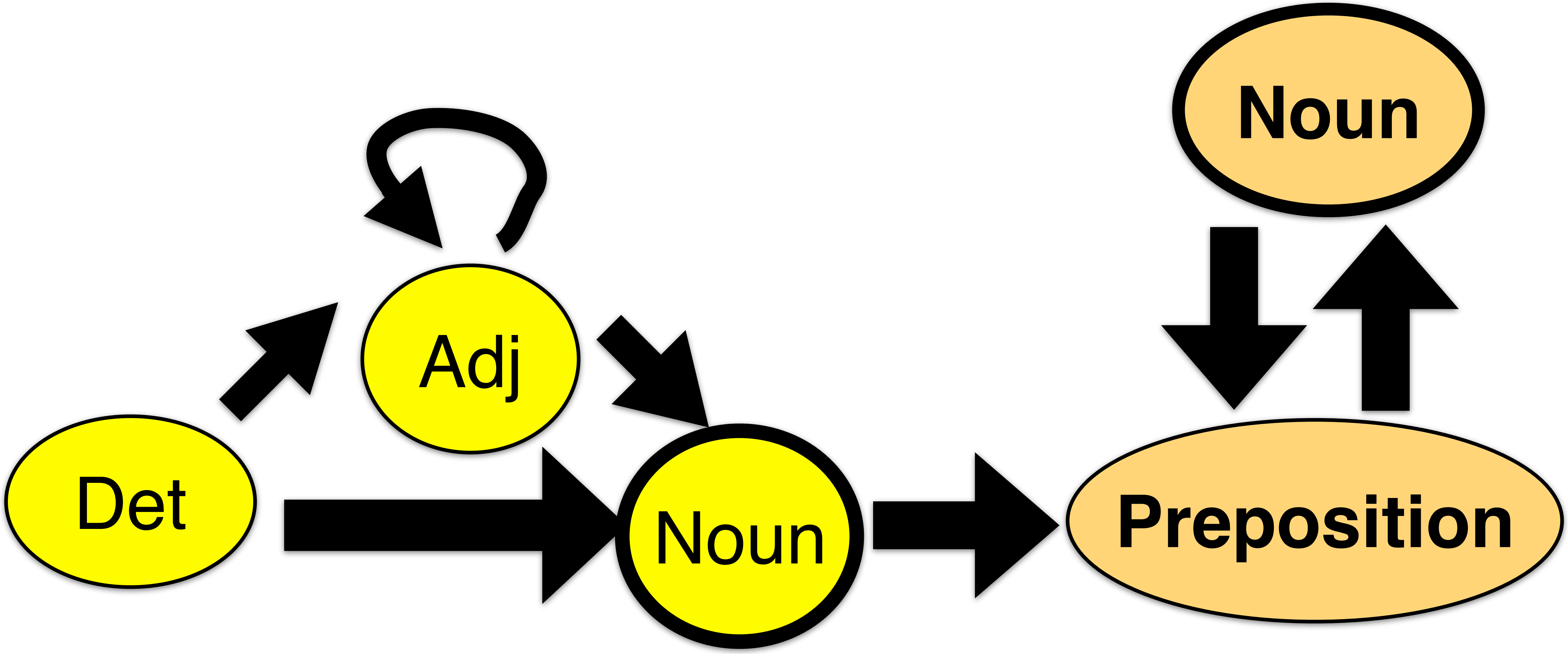
Recursion can be more complex

the ball
the ball in the garden
the ball in the garden behind the house
the ball in the garden behind the house next to the school
....

Yet another FSA



Yet another FSA



So, what do we need *grammar* for?

What does this *mean*?

the ball   in the garden behind the house

What does this *mean*?

the ball   in the garden behind the house

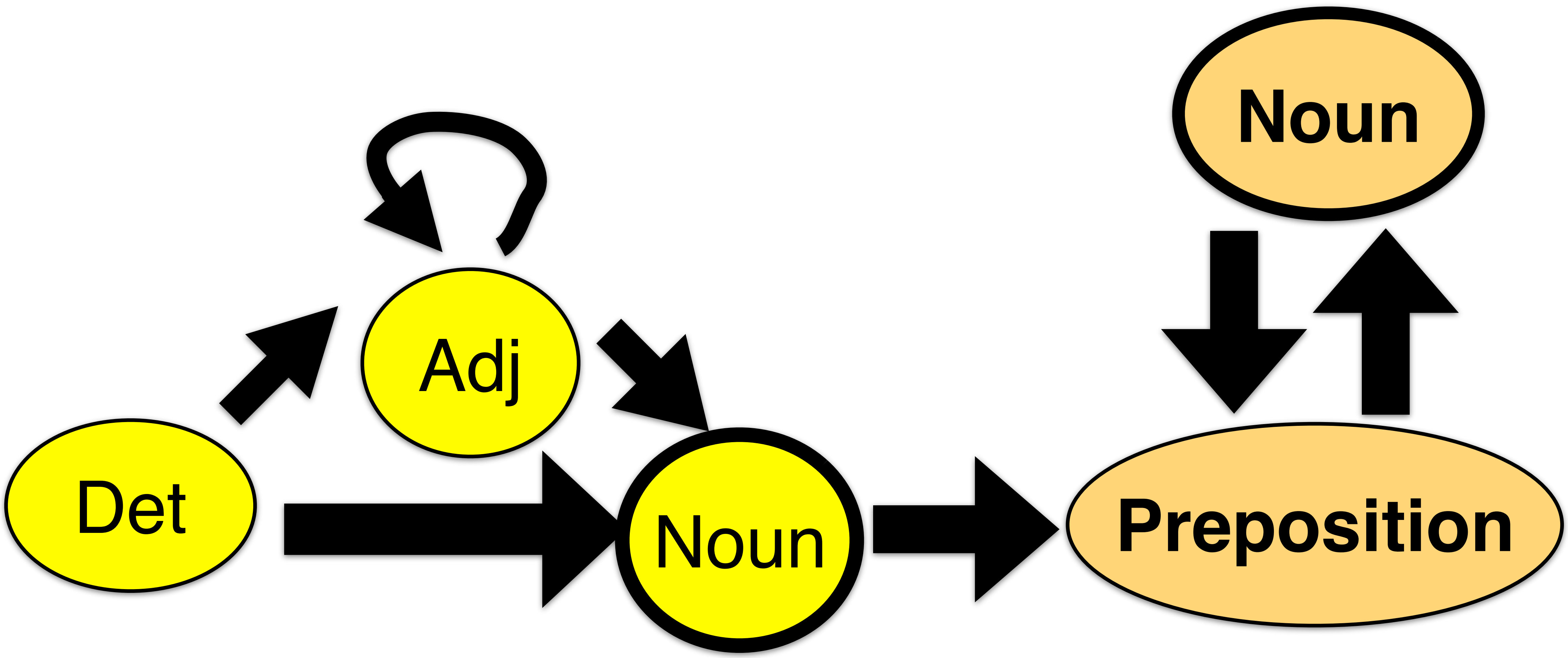
What does this *mean*?

the ball in the garden behind the house

What does this *mean*?

the ball in the garden behind the house

The FSA does not generate structure



Strong vs. weak generative capacity

- **Formal language theory:**
  - defines language as string sets
  - is only concerned with generating these strings **(weak generative capacity)**

- **Formal/Theoretical syntax (in linguistics):**
  - defines language as sets of strings with (hidden) structure
  - is also concerned with generating the right structures **(strong generative capacity)**

Context-free grammars (CFGs) capture recursion

- **Language has complex constituents**
  (*“the garden behind the house”*)
- **Syntactically, these constituents behave just like simple ones.**
  (*“behind the house”* can always be omitted)
- **CFGs define nonterminal categories to capture equivalent constituents.**

An example

N $\rightarrow$ *{ball, garden, house, sushi }*
P $\rightarrow$ *{in, behind, with}*
NP $\rightarrow$ **N**
NP $\rightarrow$ **NP PP**
PP $\rightarrow$ **P NP**

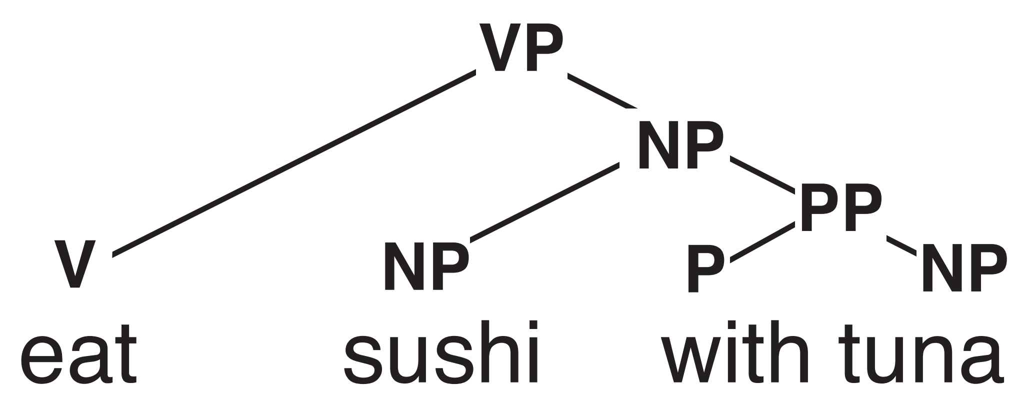
N: noun
P: preposition
NP: “noun phrase”
PP: “prepositional phrase”

Context-free grammars

- **A CFG is a 4-tuple $\langle N, \Sigma, R, S \rangle$**
  - **A set of nonterminals N**
    (e.g. N = {S, NP, VP, PP, Noun, Verb, ...})
  - **A set of terminals Σ**
    (e.g. Σ = {*I, you, he, eat, drink, sushi, ball,* })
  - **A set of rules R**
    R $\subseteq$ {$A \rightarrow \beta$ **with left-hand-side (LHS)** $A \in N$
    and right-hand-side (RHS) $\beta \in (N \cup \Sigma)^*$ }
  - **A start symbol S (sentence)**

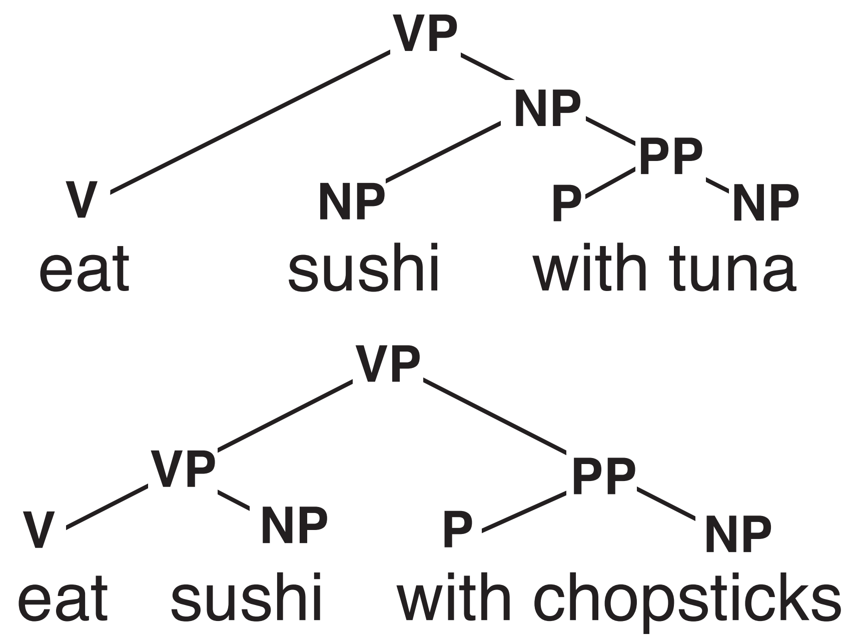
CFGs define parse trees

N → *{sushi, tuna}*
P → *{with}*
V → *{eat}*
NP → **N**
NP → **NP PP**
PP → **P NP**
VP → **V NP**



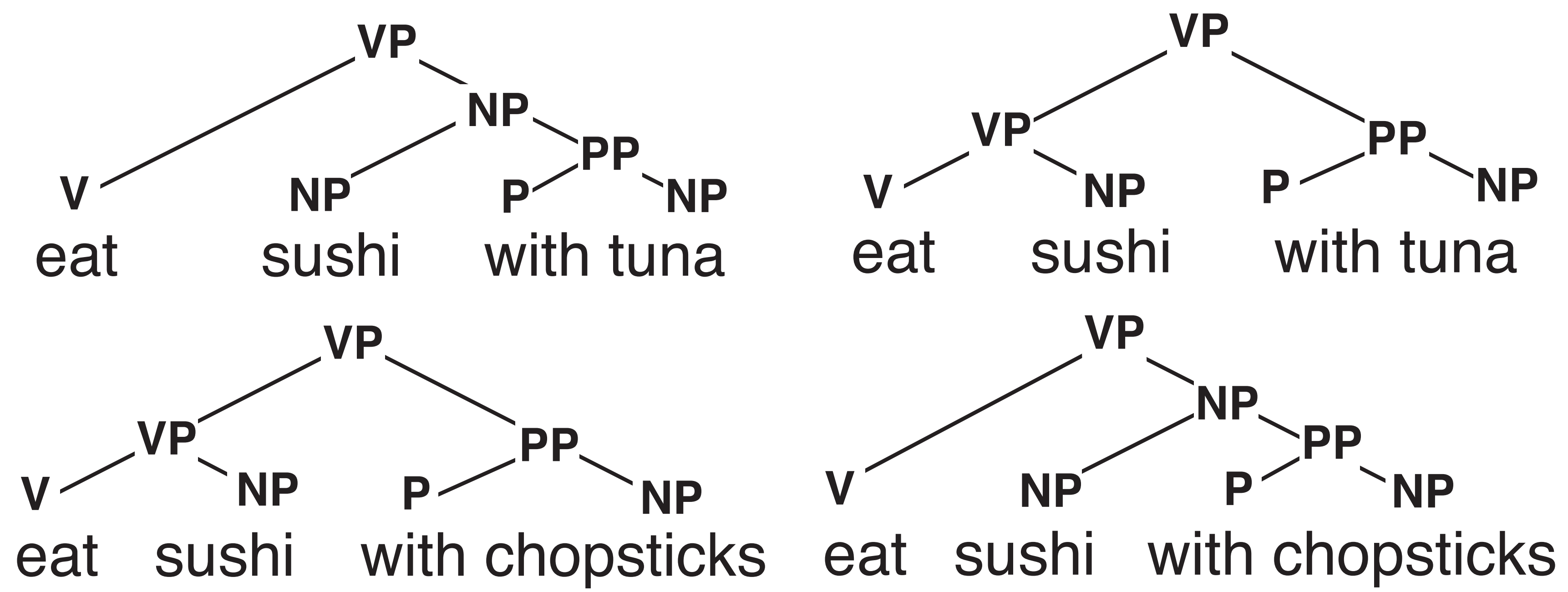
Structural ambiguity results in multiple parse trees

N → *{sushi, tuna}*
P → *{with}*
V → *{eat}*
NP → **N**
NP → **NP PP**
PP → **P NP**
VP → **V NP**
VP → **VP PP**



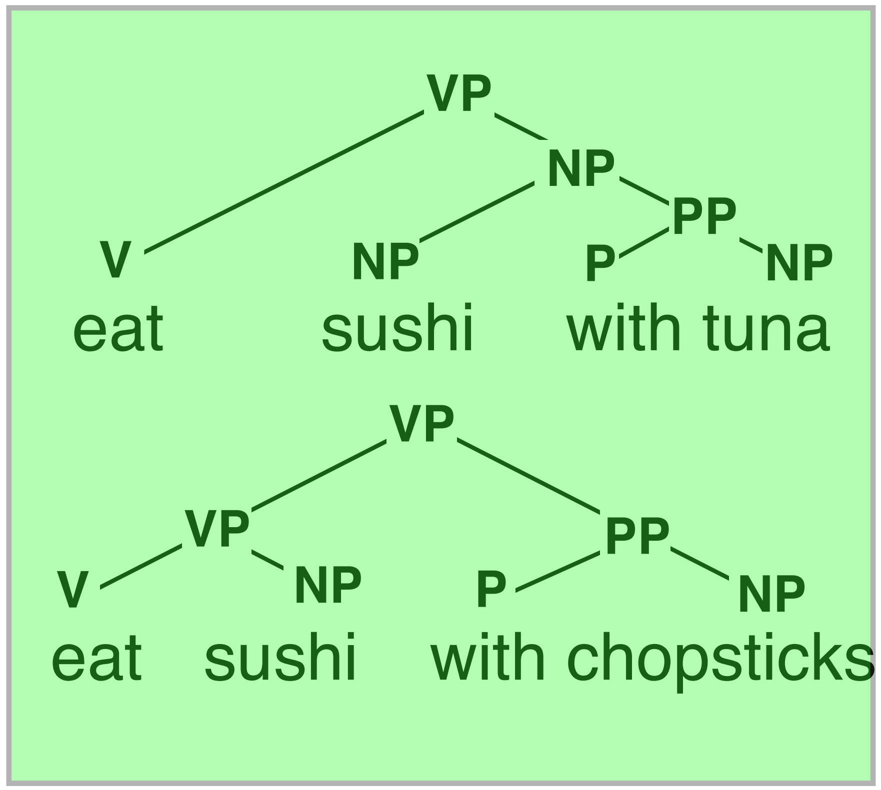
Structural ambiguity results in multiple parse trees

N → *{sushi, tuna}*
P → *{with}*
V → *{eat}*
NP → **N**
NP → **NP PP**
PP → **P NP**
VP → **V NP**
VP → **VP PP**

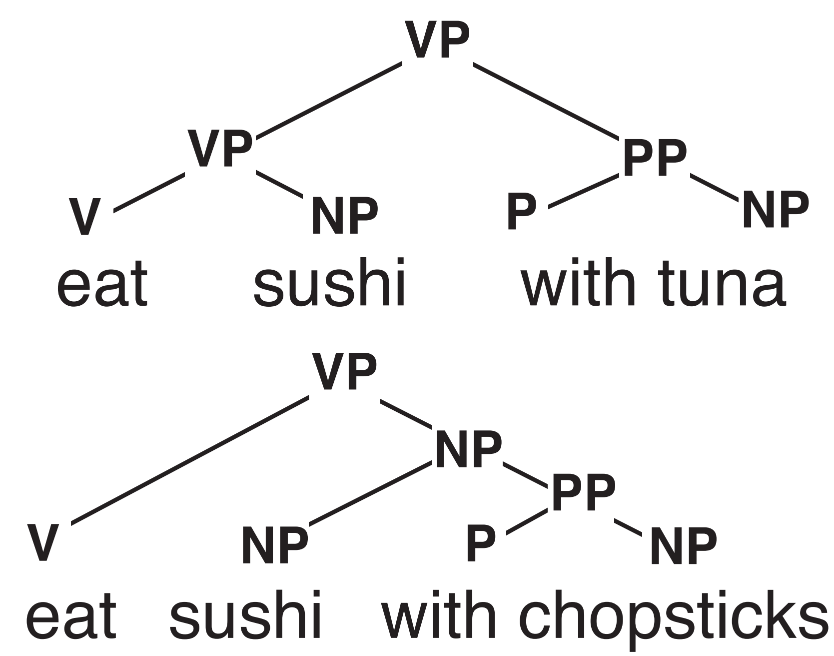


Structural ambiguity results in multiple parse trees

N → *{sushi, tuna}*
P → *{with}*
V → *{eat}*
NP → **N**
NP → **NP PP**
PP → **P NP**
VP → **V NP**
VP → **VP PP**

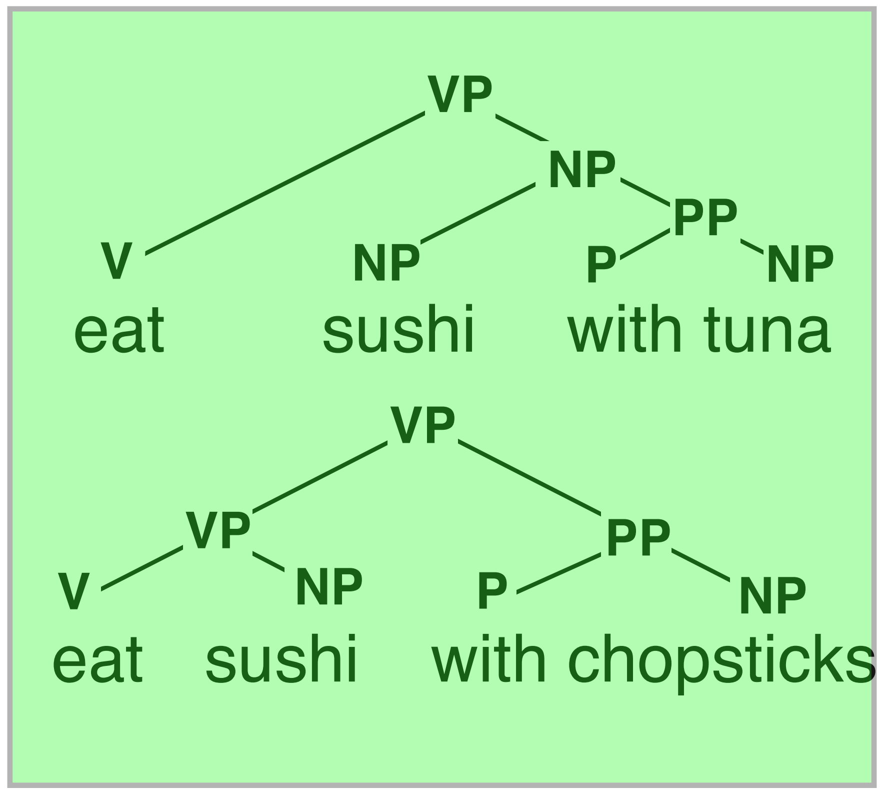


Correct Structures

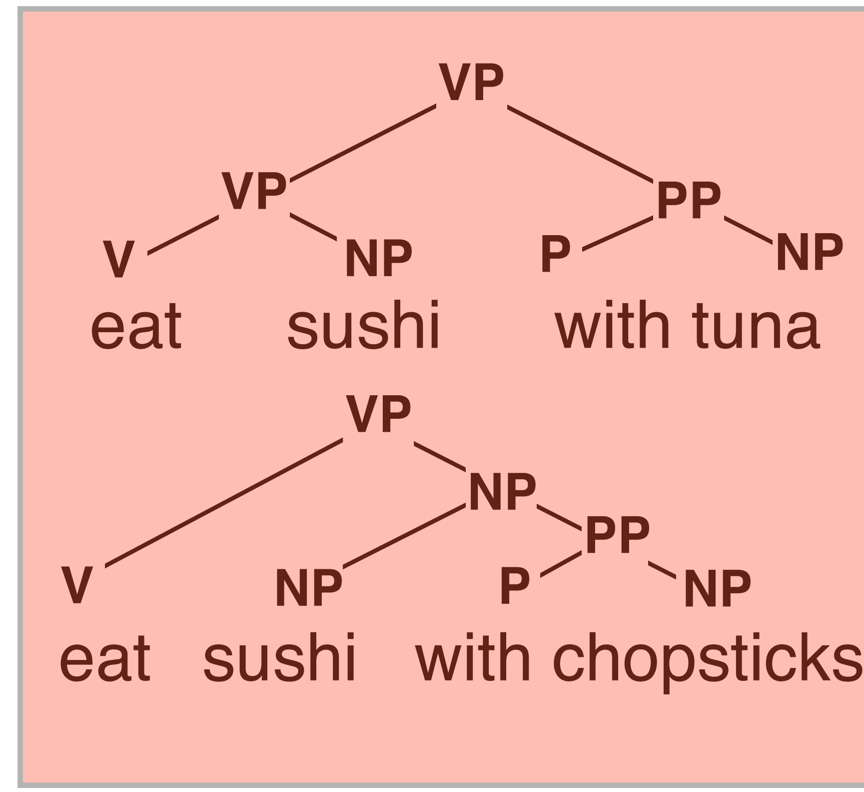


Structural ambiguity results in multiple parse trees

N → *{sushi, tuna}*
P → *{with}*
V → *{eat}*
NP → N
NP → NP PP
PP → P NP
VP → V NP
VP → VP PP



Correct Structures



Incorrect Structures

A grammar for a fragment of English

Is string α a constituent?

He talks *[in class]*.

Is string α a constituent?

He talks *[in class]*.

- **Substitution test:**
  Can α be replaced by a single word?
  He talks [there].
- **Movement test:**
  Can α be moved to in the sentence?
  [In class], he talks.
- **Answer test:**
  Can α be the answer to a question?
  Where does he talk? - [In class].

Noun phrases (NPs)

Simple NPs:

[He] sleeps. **(pronoun)**

[John] sleeps. **(proper name)**

[A student] sleeps. **(determiner + noun)**

Complex NPs:

[A tall student] sleeps. **(det + adj + noun)**

[The student in the back] sleeps. **(NP + PP)**

[The student who likes MTV] sleeps. **(NP + Relative Clause)**

The NP fragment

NP → Pronoun
NP → ProperName
NP → Det Noun

Det → *{a, the, every}*
Pronoun → *{he, she,...}*
ProperName → *{John, Mary,...}*
Noun → AdjP Noun
Noun → N
NP → NP PP
NP → NP RelClause

Adjective phrases and Prepositional Phrases

AdjP → Adj
AdjP → Adv AdjP
Adj → *{big, small, red,...}*
Adv → *{very, really,...}*

PP → P NP
P → *{with, in, above,...}*

The Verb Phrase (VP)

He [eats].
He [eats sushi].
He [gives John sushi].
He [eats sushi with chopsticks].

VP → V
VP → V NP
VP → V NP PP
VP → VP PP

V → *{eats, sleeps gives,...}*

VPs redefined

He [eats].
He [eats sushi].
He [gives John sushi].
He [eats sushi with chopsticks].

VP → V_Intrans
VP → V_trans NP
VP → V_ditrans NP NP
VP → VP PP
V_intrans→ *{eats, sleeps}*
V_trans→ *{eats}*
V_trans→ *{gives}*

Sentences

[He eats sushi].
[Sometimes, he eats sushi].
[In Japan, he eats sushi].

S → NP VP
S → AdvP S
S → PP S

He says [he eats sushi].
VP → V_comp S
V_comp → *{says, think, believes}*

Sentences redefined

[He eats sushi]. ✔
[I eats sushi].* **???
[They eats sushi].* **???

S → NP.3sg VP.3sg
S → NP.1sg VP.1sg
S → NP.3pl VP.3pl

We need features to capture agreement:
(number, person, case,...)

More on verbs

Tense:

He [eats]. **Present tense**

He [ate]. **Past tense**

He [has eaten]. **Present perfect tense**

He [will eat]. **Future tense**

Voice:

He [is/was eaten]. **Passive voice**

Aspect:

He [is/was eating]. **Progressive**

Mood:

He [could eat]. **Conditional**

Different kinds of verbs

Main verbs (eat,...) and their forms:

*He [**eats**].* **Present tense form**
*He [**ate**].* **Past tense form**
*He [has **eaten**].* **Past participle**
*He [is/was **eating**].* **Present participle**
*He [will **eat**].* **(bare) infinitive**

Auxiliary verbs (for tense and voice):

be (am, are, is, was, will, would...)
have (has, had, ...)

Modals:

must, can, should, ...

Morphology and syntax

- **English has very simple morphology:**
  - "eat": infinitive, 1&2 pers sg/pl present , 3pers pl present
- **Many languages (German, Latin, Russian, Finnish) have more complex morphology:**
  - "isst": 2 pers sg present tense
- **In such languages, word order is a lot freer than in English**