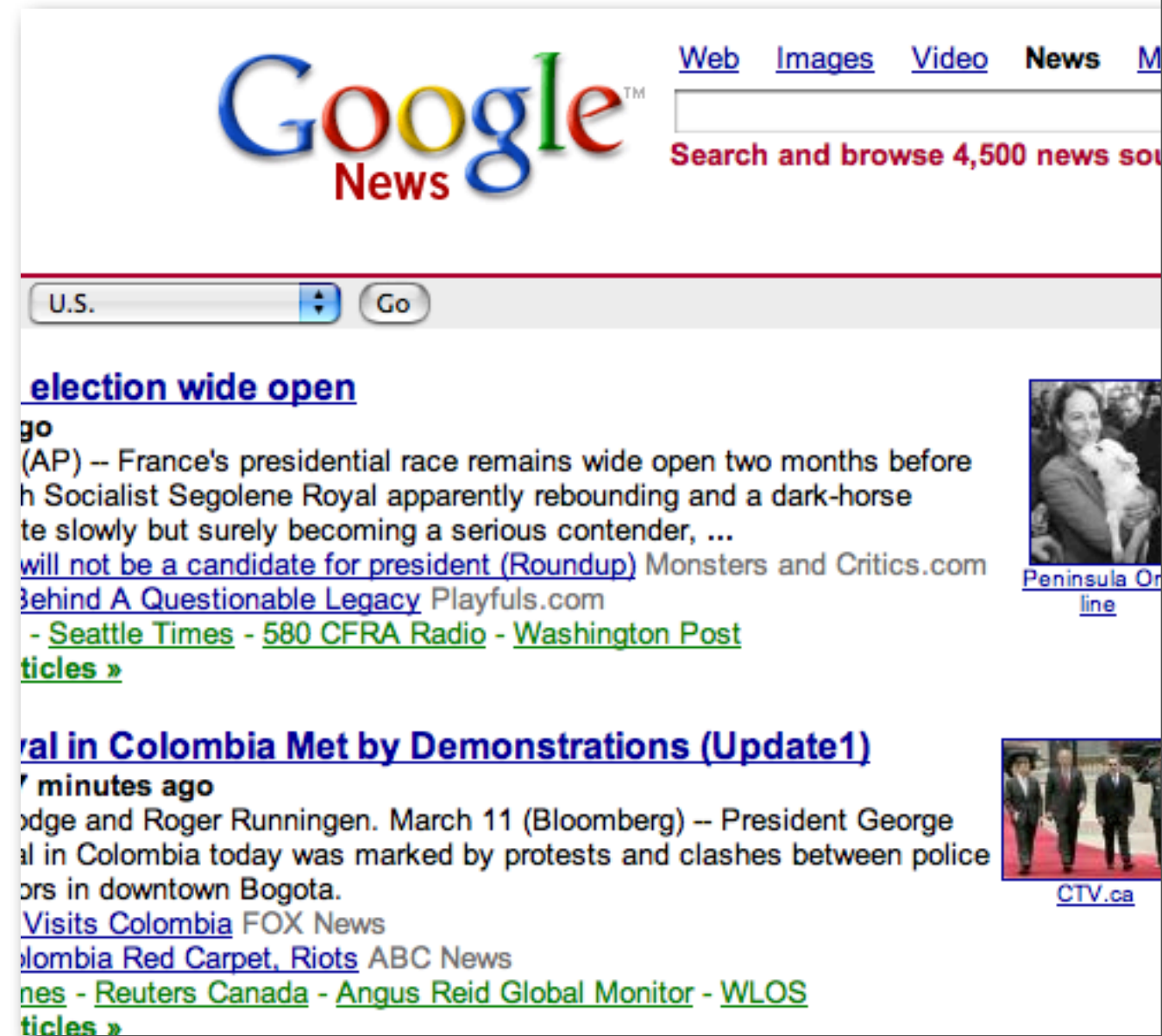


CS 598

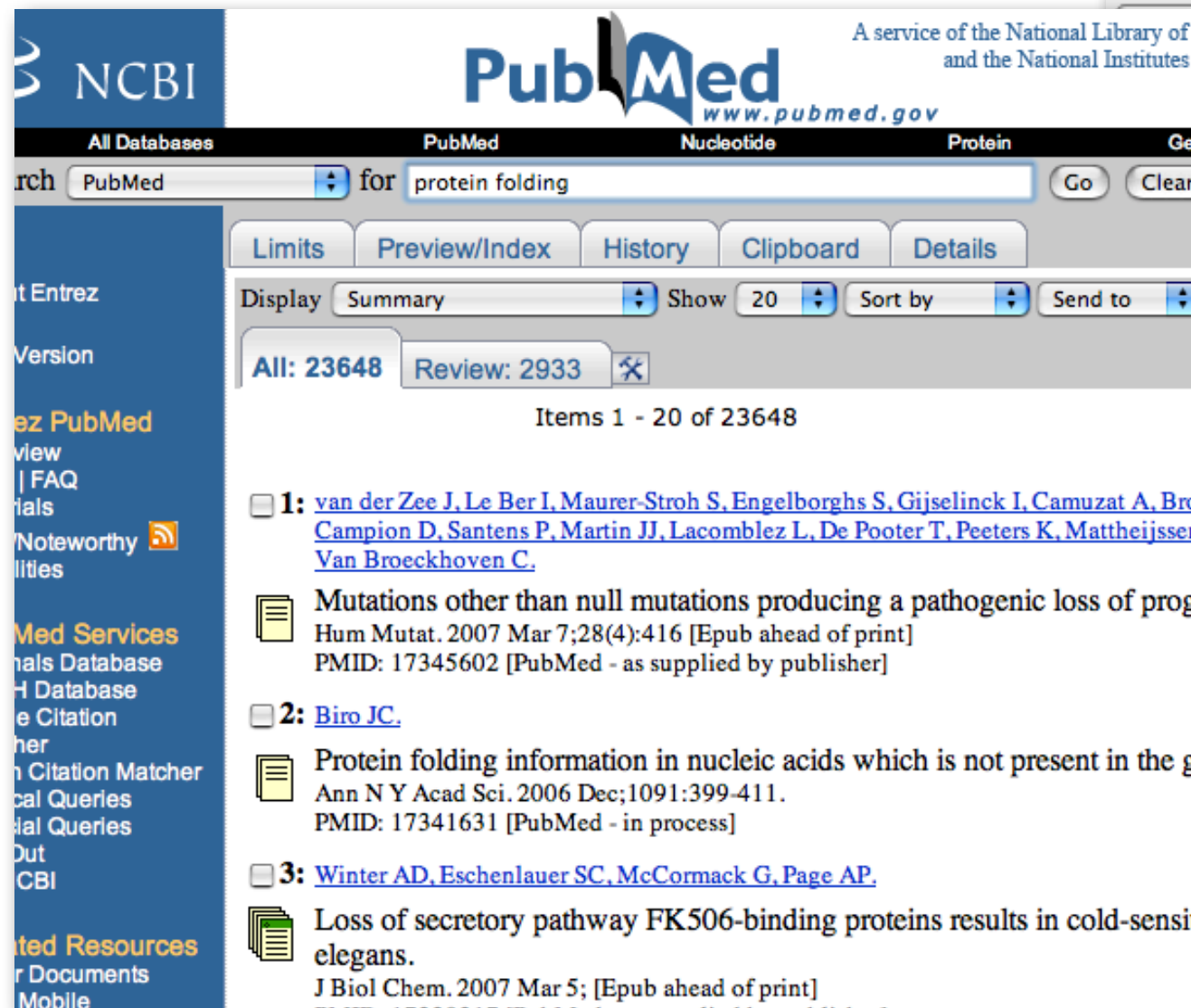
**Natural Language
Processing**

Natural language is everywhere

Natural language is everywhere



Natural language is everywhere



Presidential race wide open

France's presidential race remains wide open two months before the first round of voting. Socialist Segolene Royal apparently rebounding and a dark-horse candidate but surely becoming a serious contender, ...

[be a candidate for president \(Roundup\)](#) Monsters and Critics.com

[A Questionable Legacy](#) Playfuls.com

[The New York Times](#) - [580 CFRA Radio](#) - [Washington Post](#)



[Peninsula Online](#)

Colombia Met by Demonstrations (Update1)

... days ago

and Roger Runnigen. March 11 (Bloomberg) -- President George W. Bush's visit to Colombia today was marked by protests and clashes between police and demonstrators in downtown Bogota.

[Colombia](#) FOX News

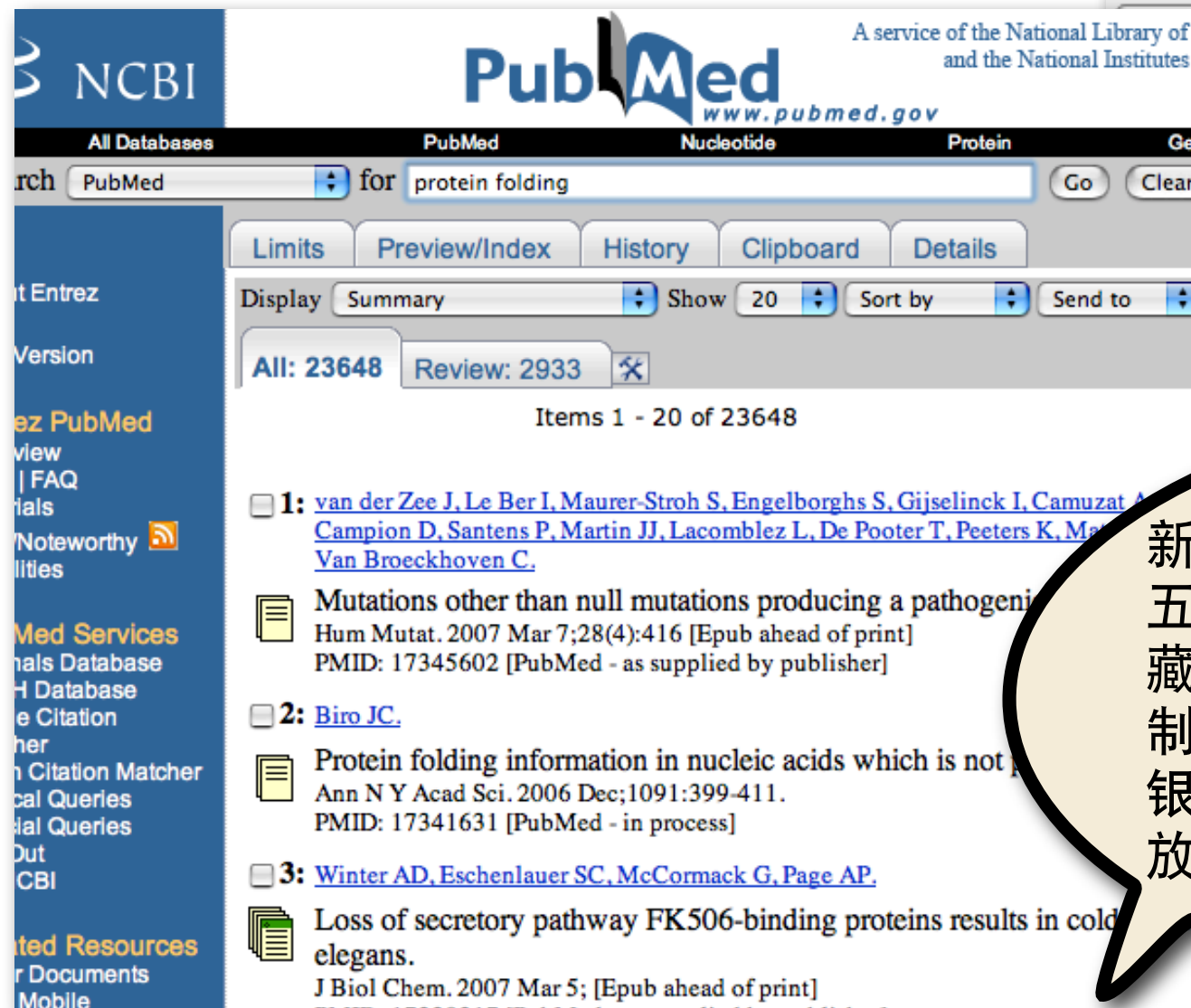
[A Red Carpet, Riots](#) ABC News

[Reuters Canada](#) - [Angus Reid Global Monitor](#) - [WLOS](#)



[CTV.ca](#)

Natural language is everywhere



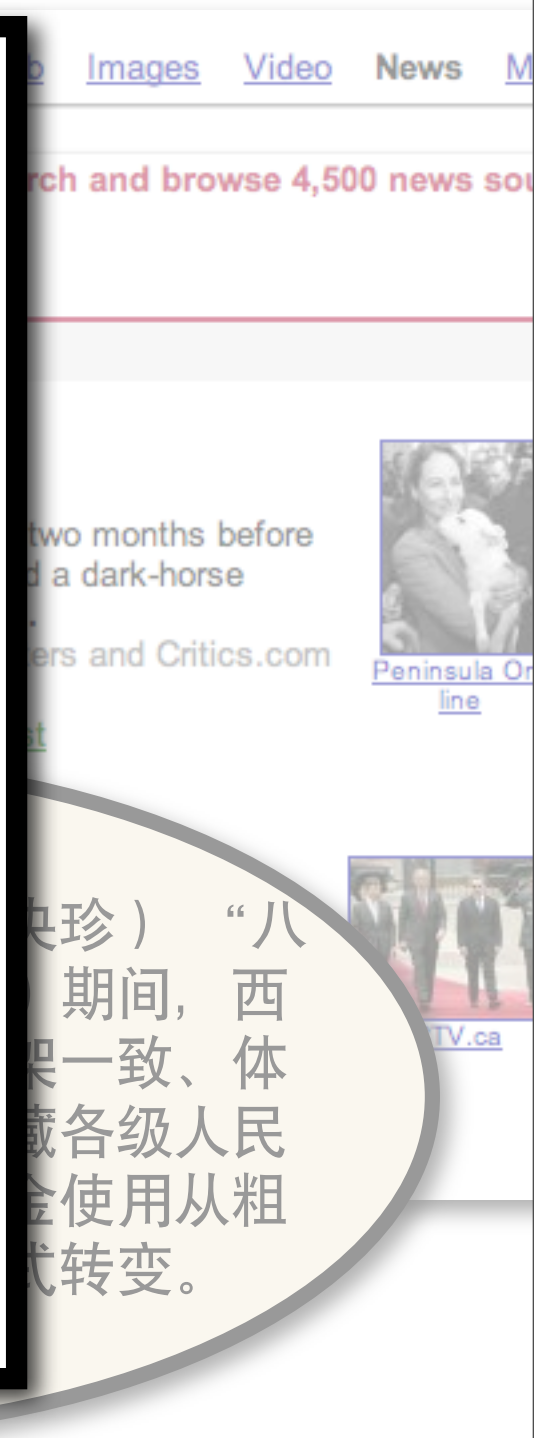
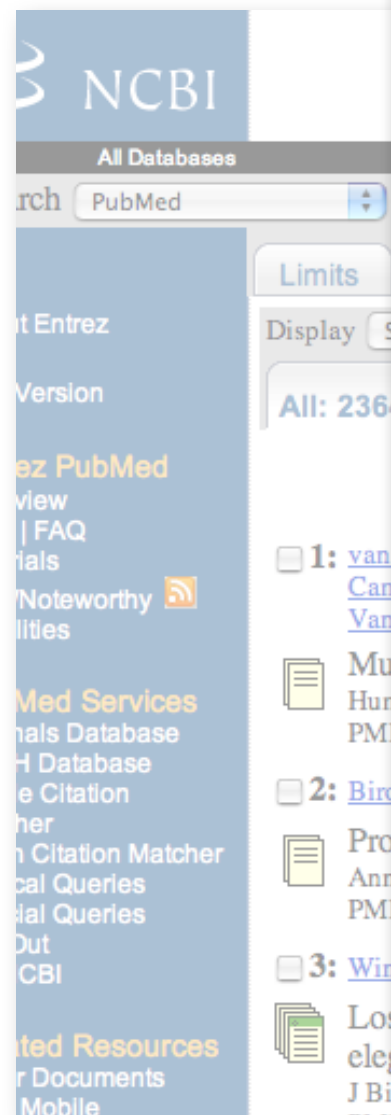
新华社拉萨二月二日电（记者央珍）“八五”（一九九一至一九九五年）期间，西藏金融体制改革坚持与全国框架一致、体制衔接的方针，顺利完成了西藏各级人民银行的分设工作，实现信贷资金使用从粗放型经营方式向集约型经营方式转变。

Natural language is everywhere

NLP applications:
Information extraction
(news, scientific papers)

Machine translation

Dialog systems
(phone, robots)



Different ways of studying language

- *How does language work?*
(core linguistics)
- *How do people learn and process language?*
(psycholinguistics)
- *Where in the brain is language located?*
(neurolinguistics)
- *How do languages change over time?*
(historical linguistics)
- *How does language express identity/social status?*
(sociolinguistics)
- *How can you teach foreign languages?*
(applied linguistics)

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)

Computational Linguistics/ Natural Language Processing

- ***Can we build computational systems that process language?***
- ***Process:***
translate, understand, summarize, generate,....
- **Text-based: Requires (at least) morphology, syntax, semantics (pragmatics is hard)**
- **Speech-based: also phonetics/phonology**

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?

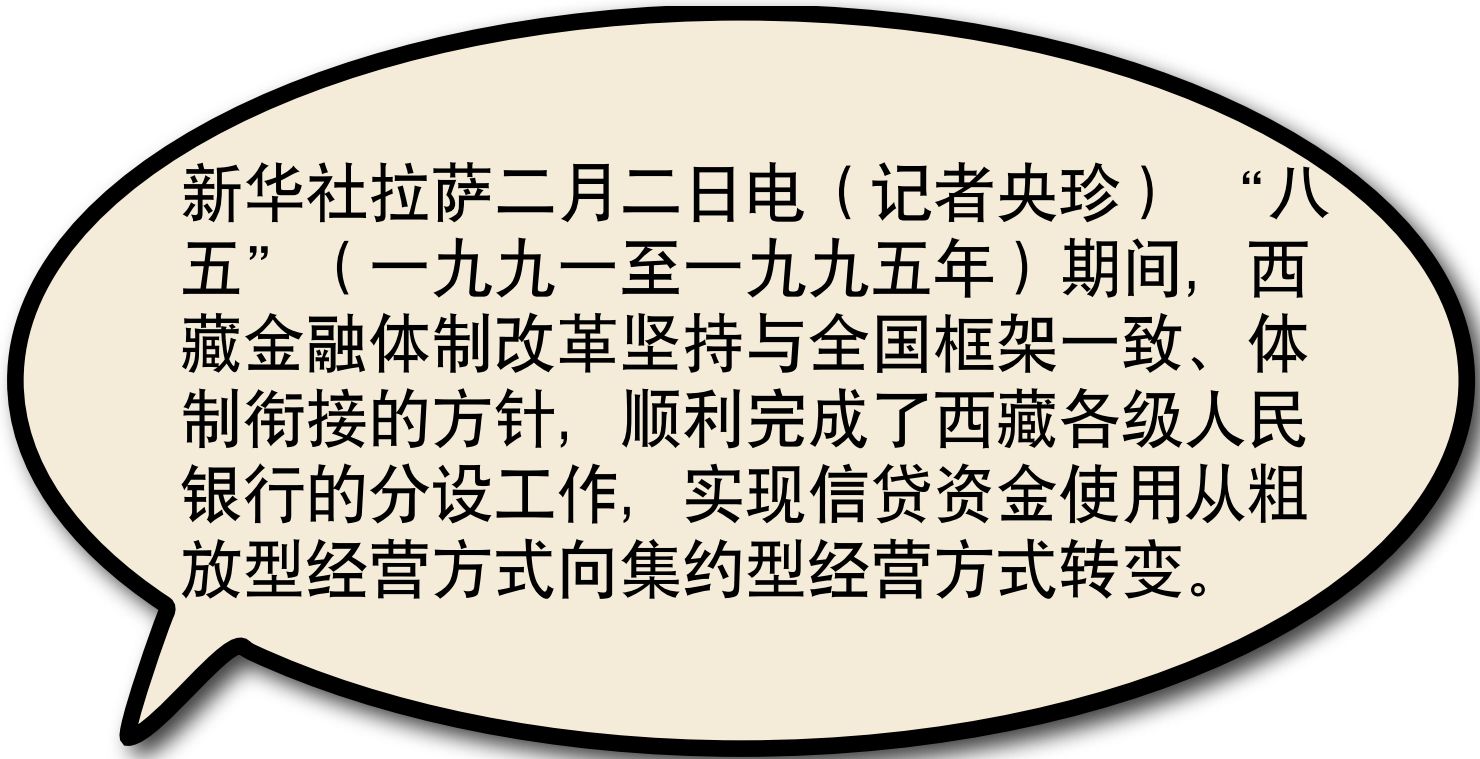
Research trends in NLP

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!**

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 tuna.

I eat sushi with tuna.

I eat sushi with chopsticks.

I eat sushi with tuna.

I eat sushi with chopsticks.

I eat **sushi with tuna.**

I **eat** **sushi** **with chopsticks.**

Language is ambiguous.

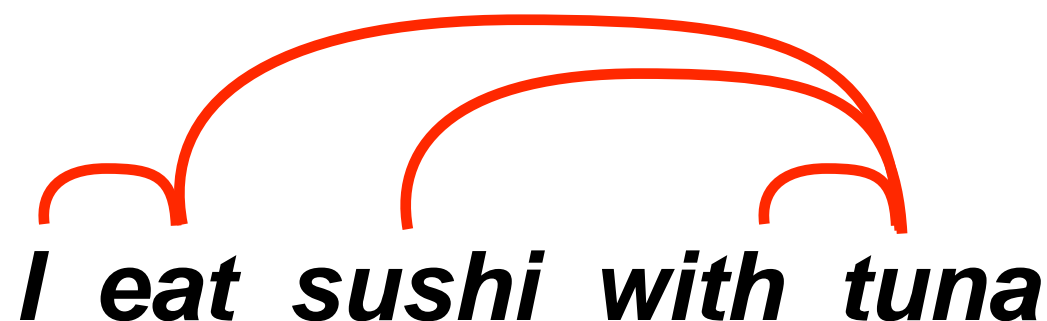
Statistical models:

What is the most likely structure?

We need a probability model.

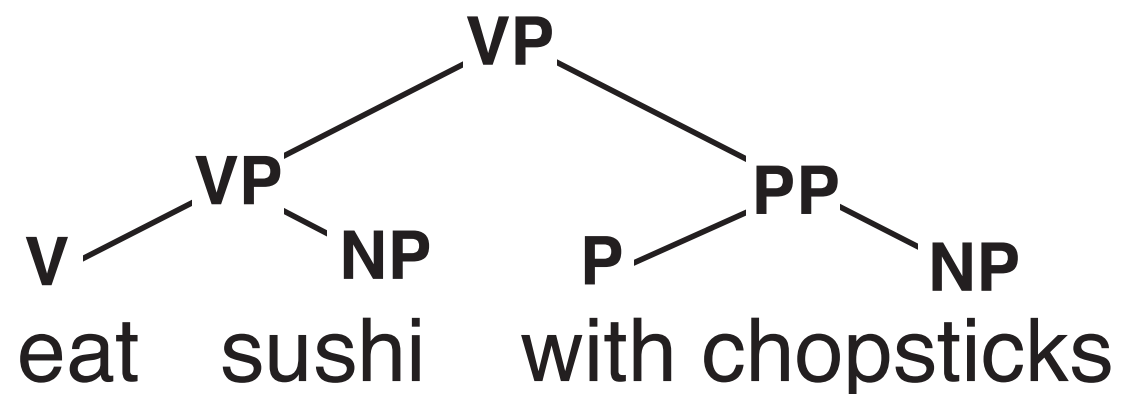
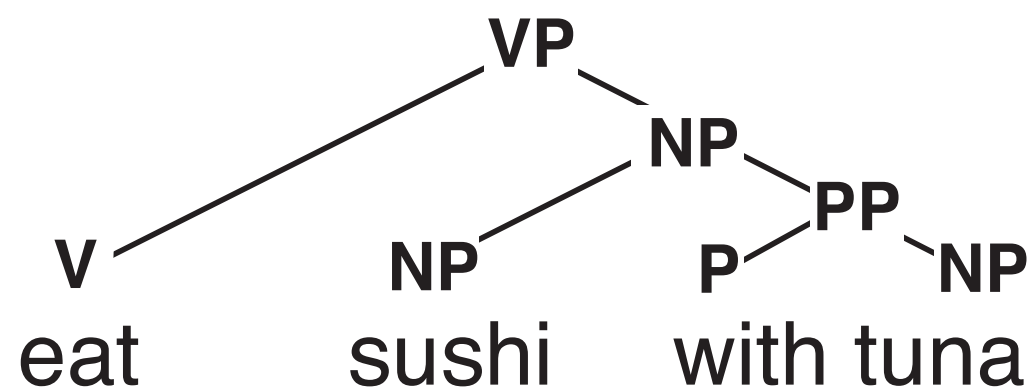
What is the structure of a sentence?

- **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

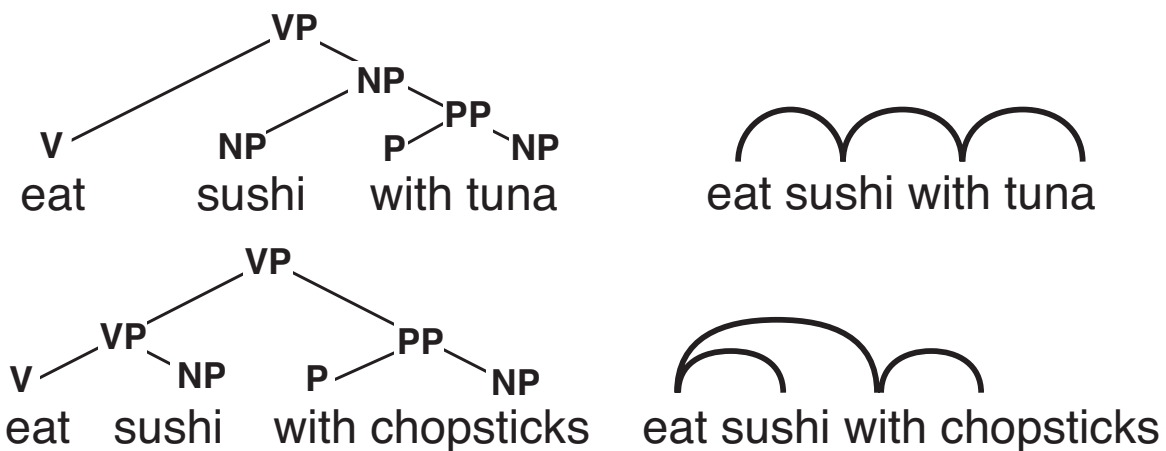


Dependency trees

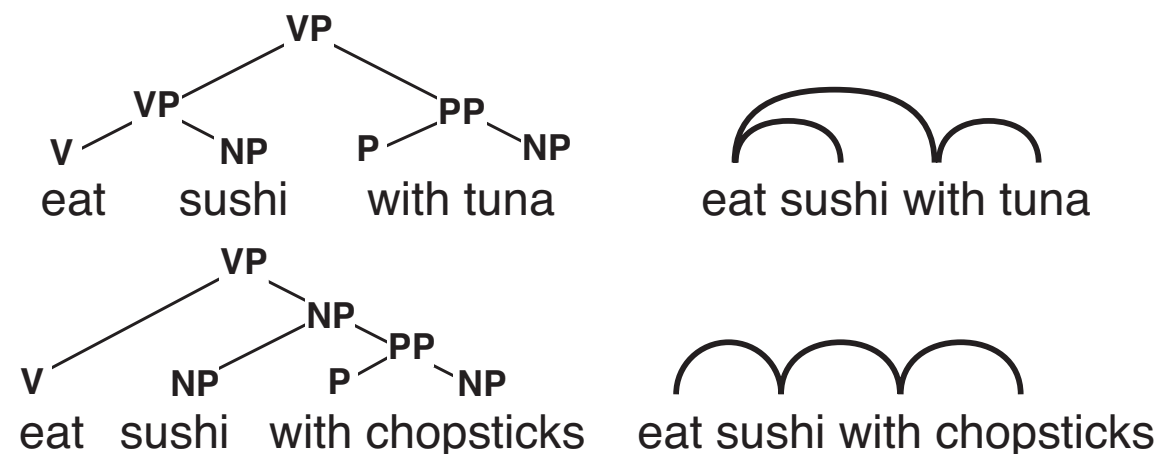


Structure (Syntax) corresponds to Meaning (Semantics)

Correct analysis



Incorrect analysis



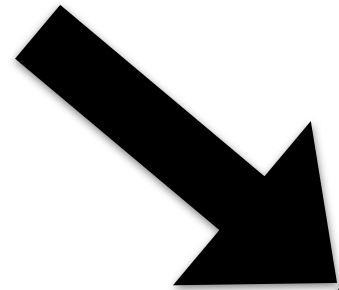
The goal of formal syntax:
***Can we define a program that
generates all English sentences?***

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 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?

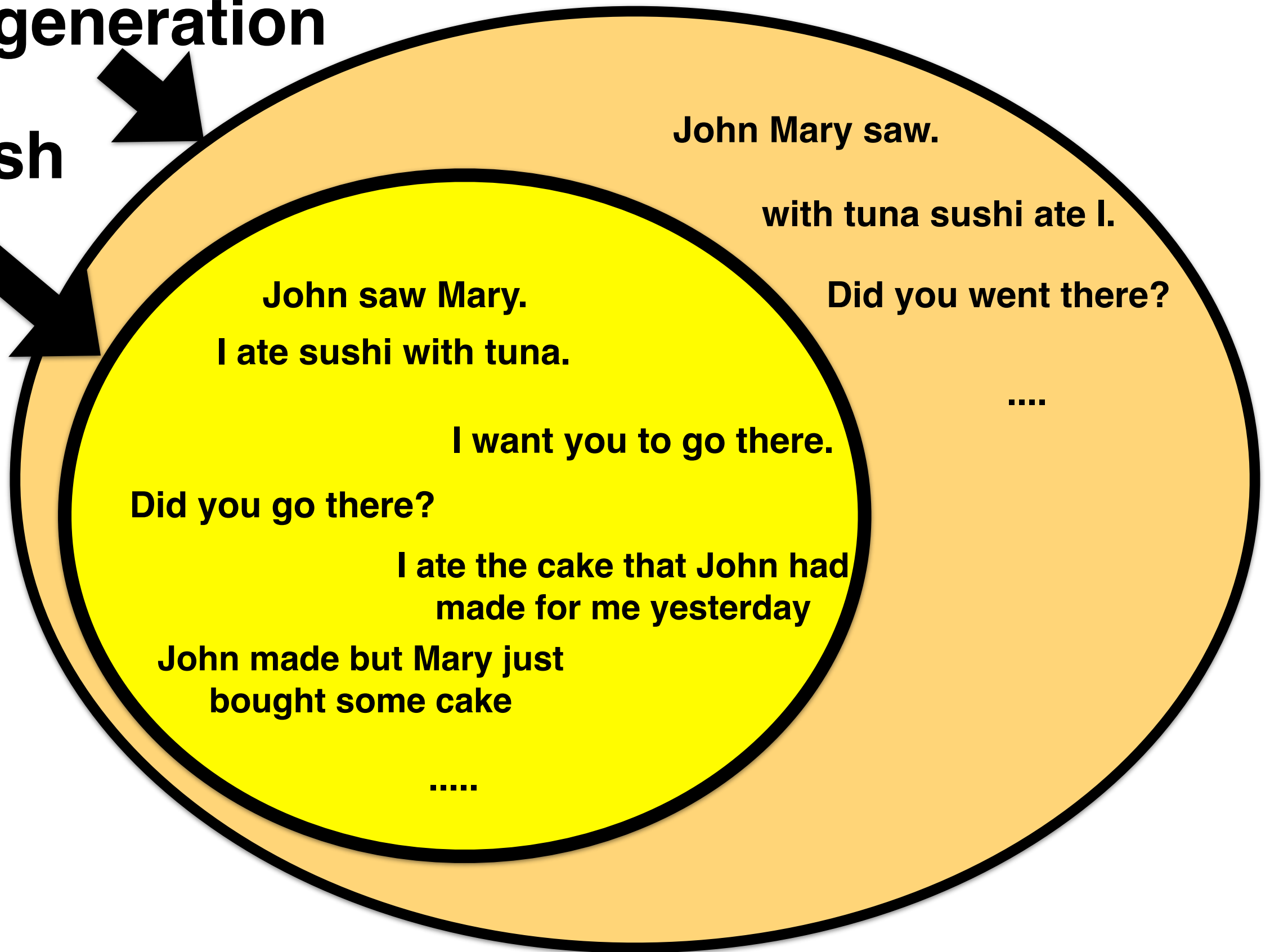
I ate the cake that John had
made for me yesterday

John made but Mary just
bought some cake

.....

Overgeneration

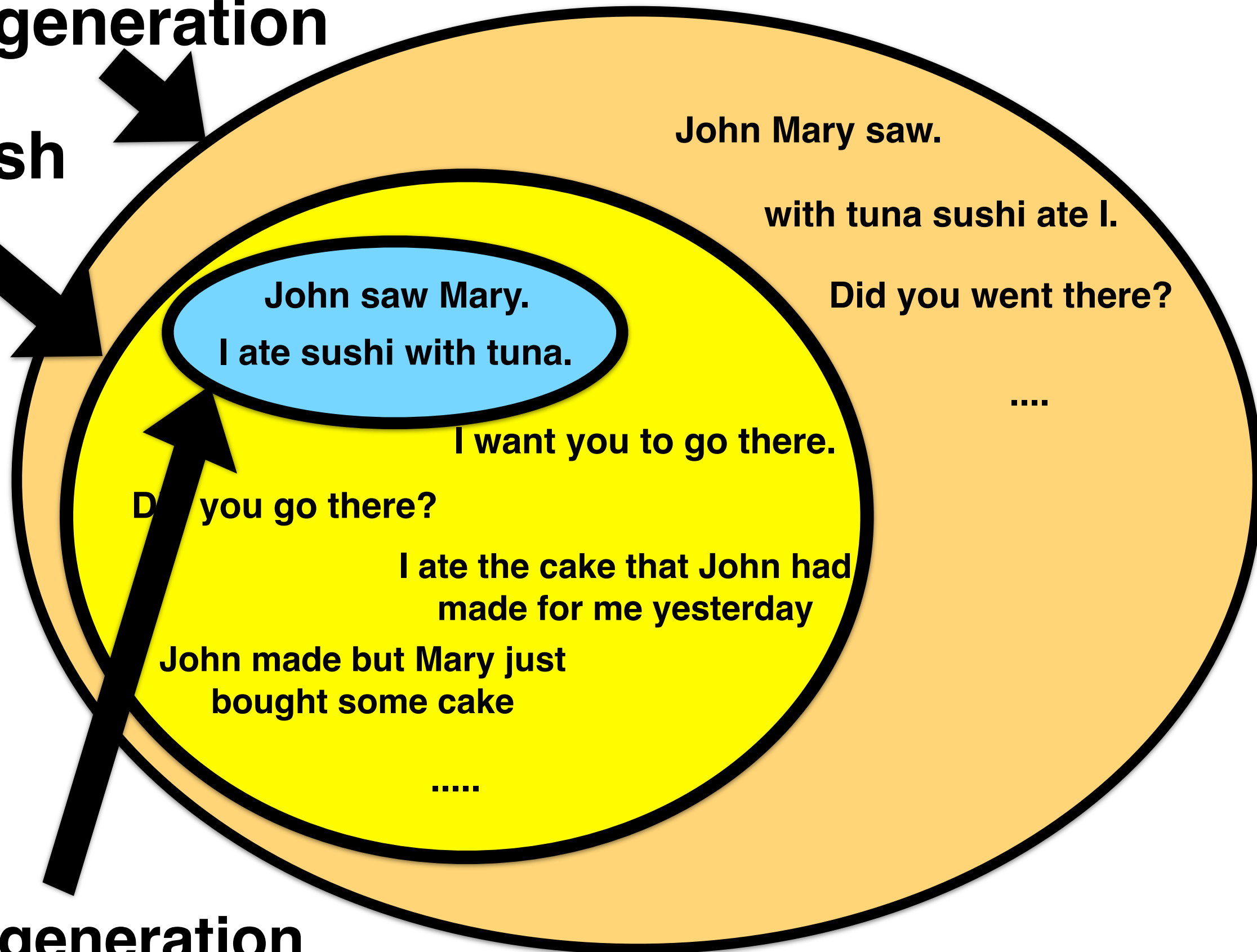
English



Overgeneration

English

Undergeneration



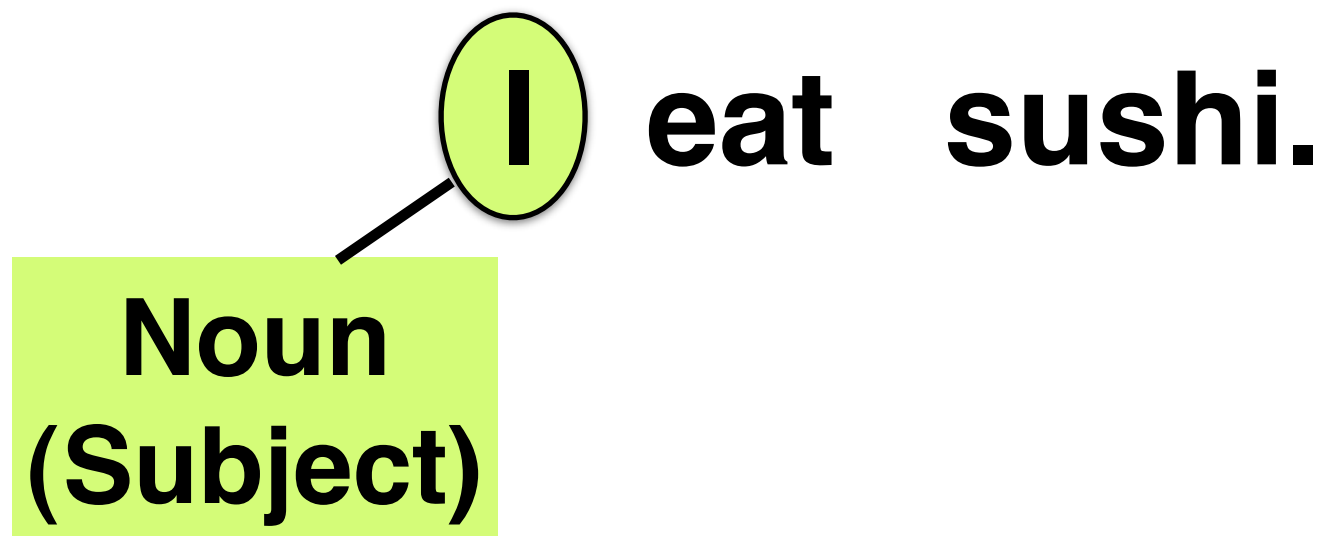
Basic word classes (parts of speech)

- **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*

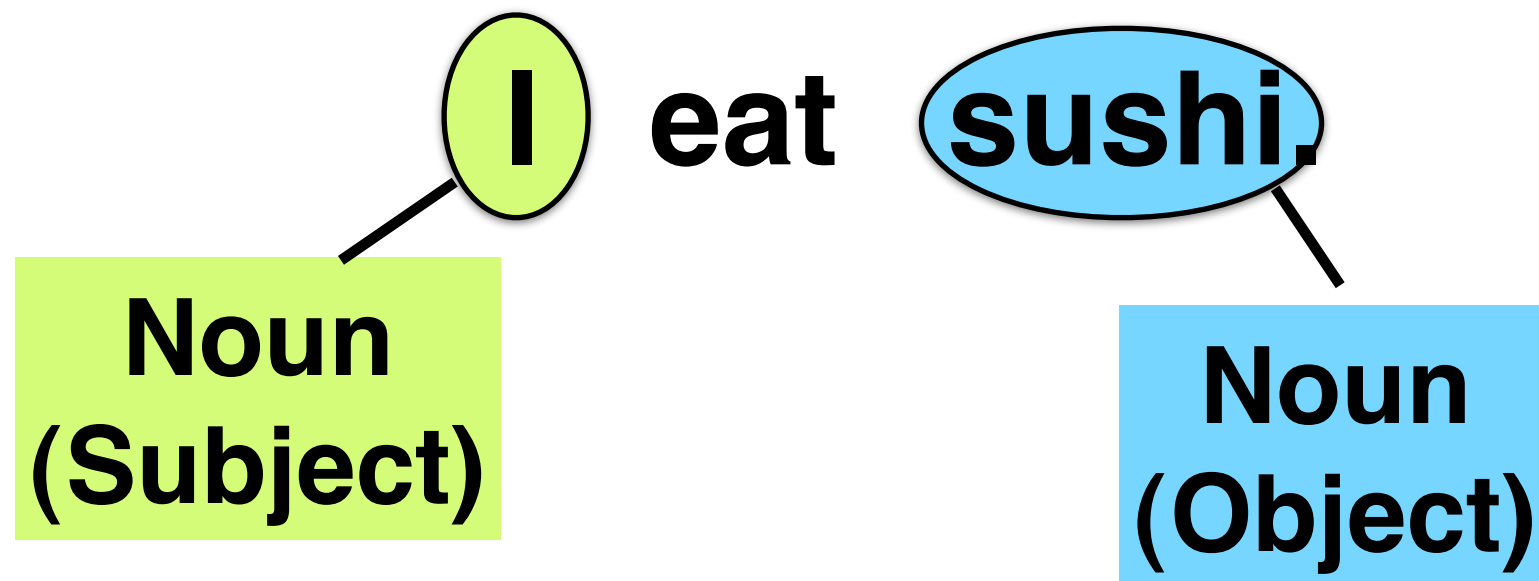
Basic sentence structure

I eat sushi.

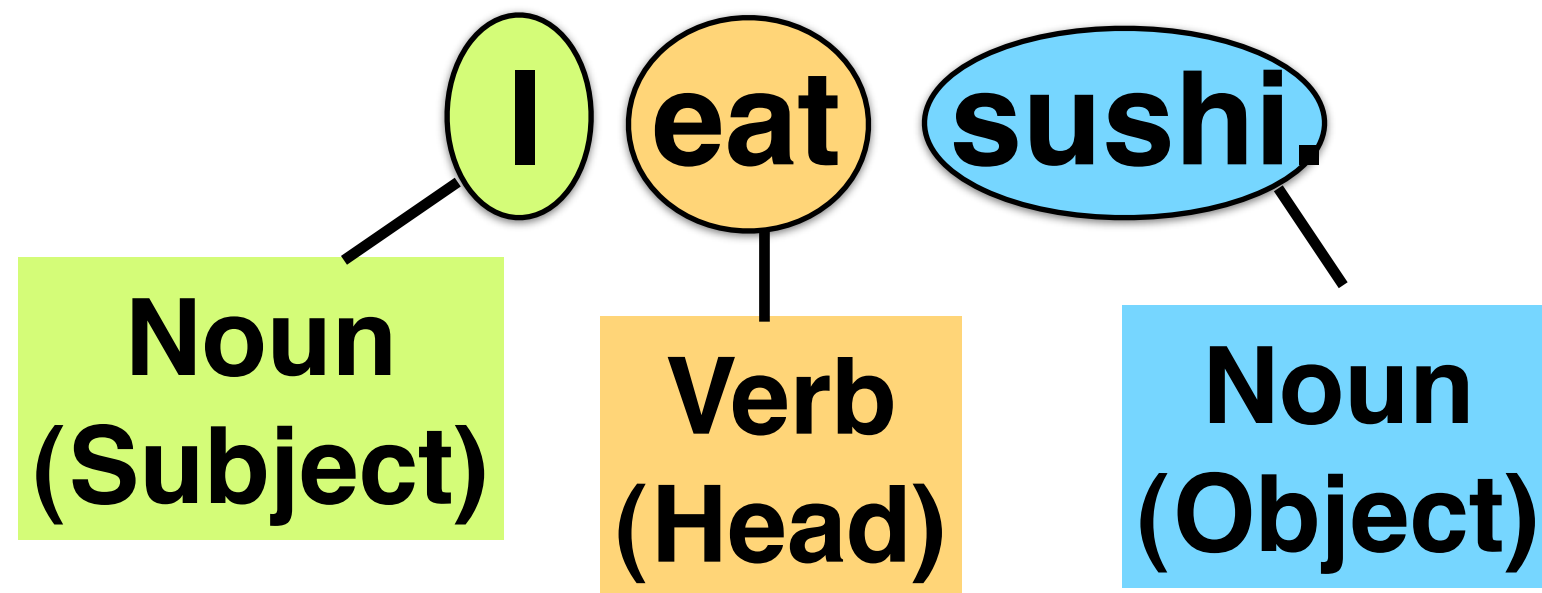
Basic sentence structure



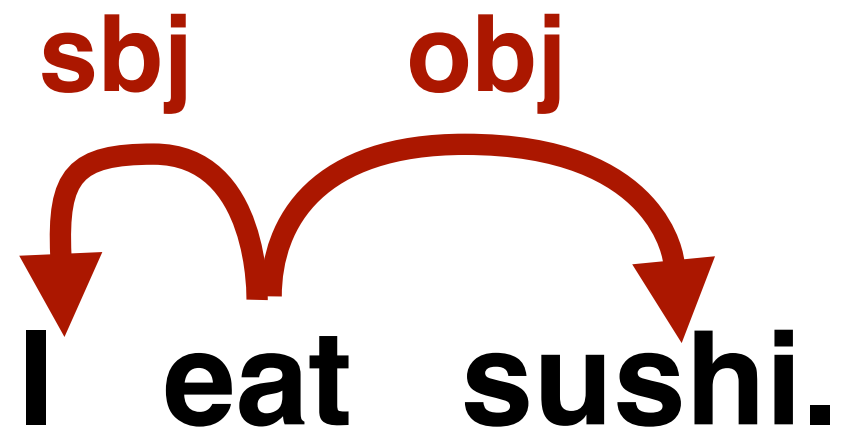
Basic sentence structure



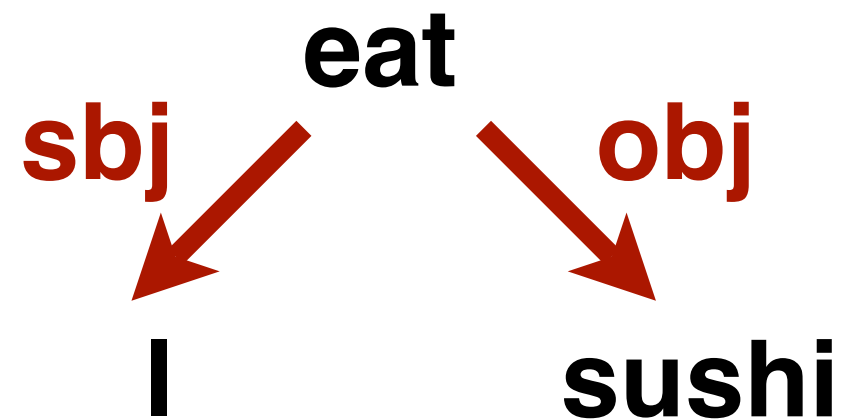
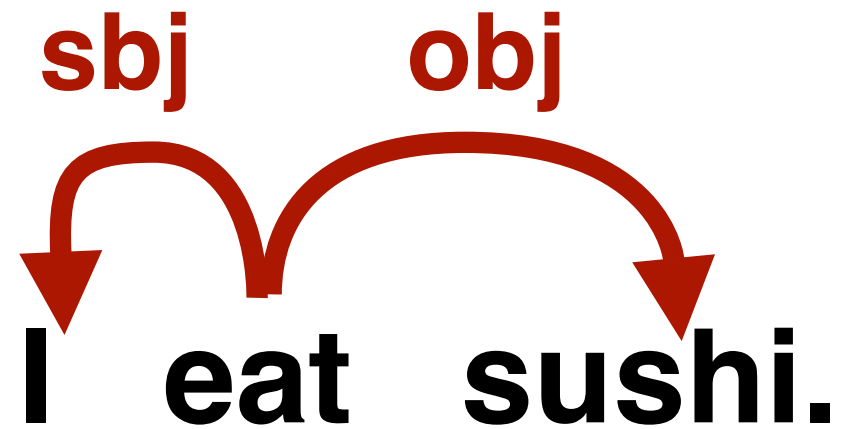
Basic sentence structure



As a dependency tree



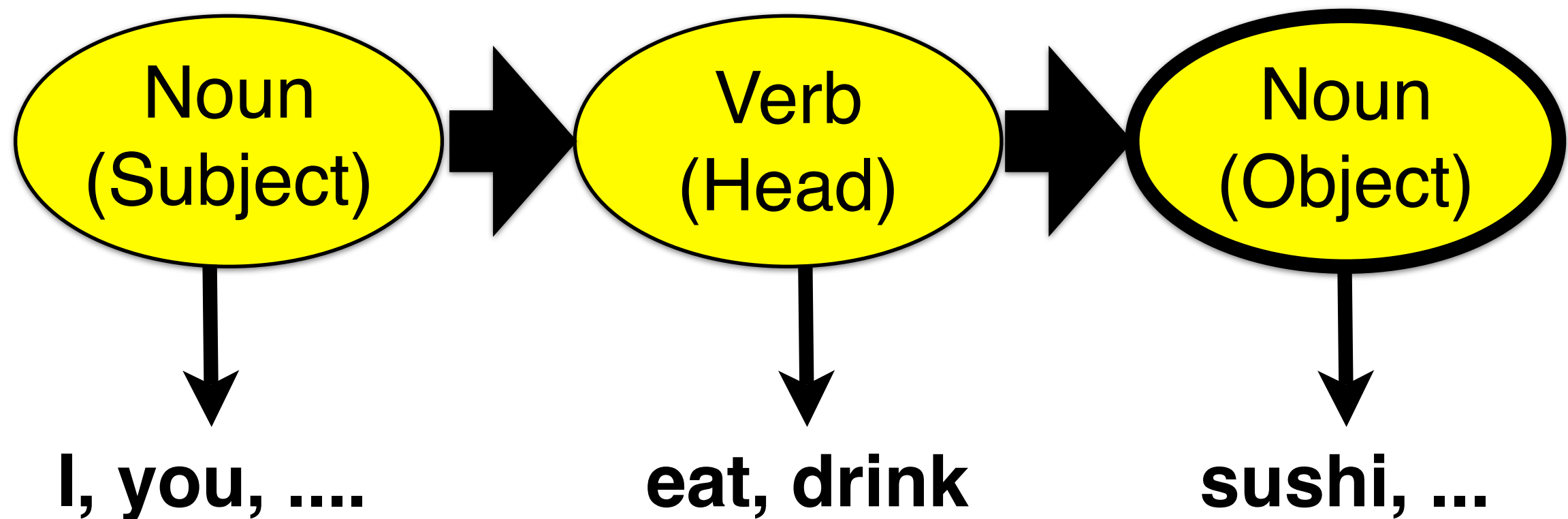
As a dependency tree



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.

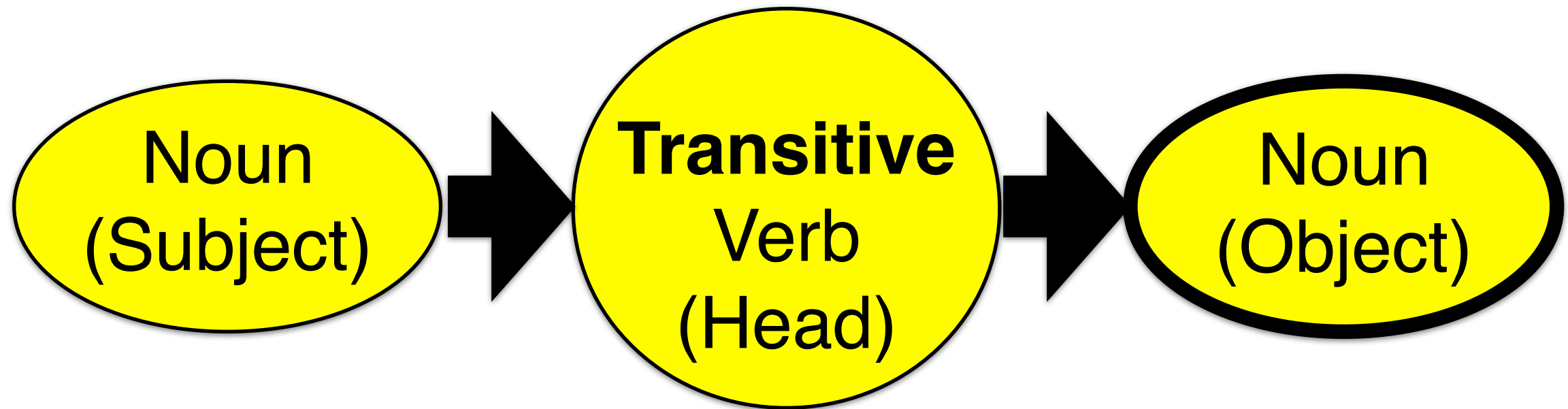
Transitive verbs (eat) take also one (direct) object.

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

....

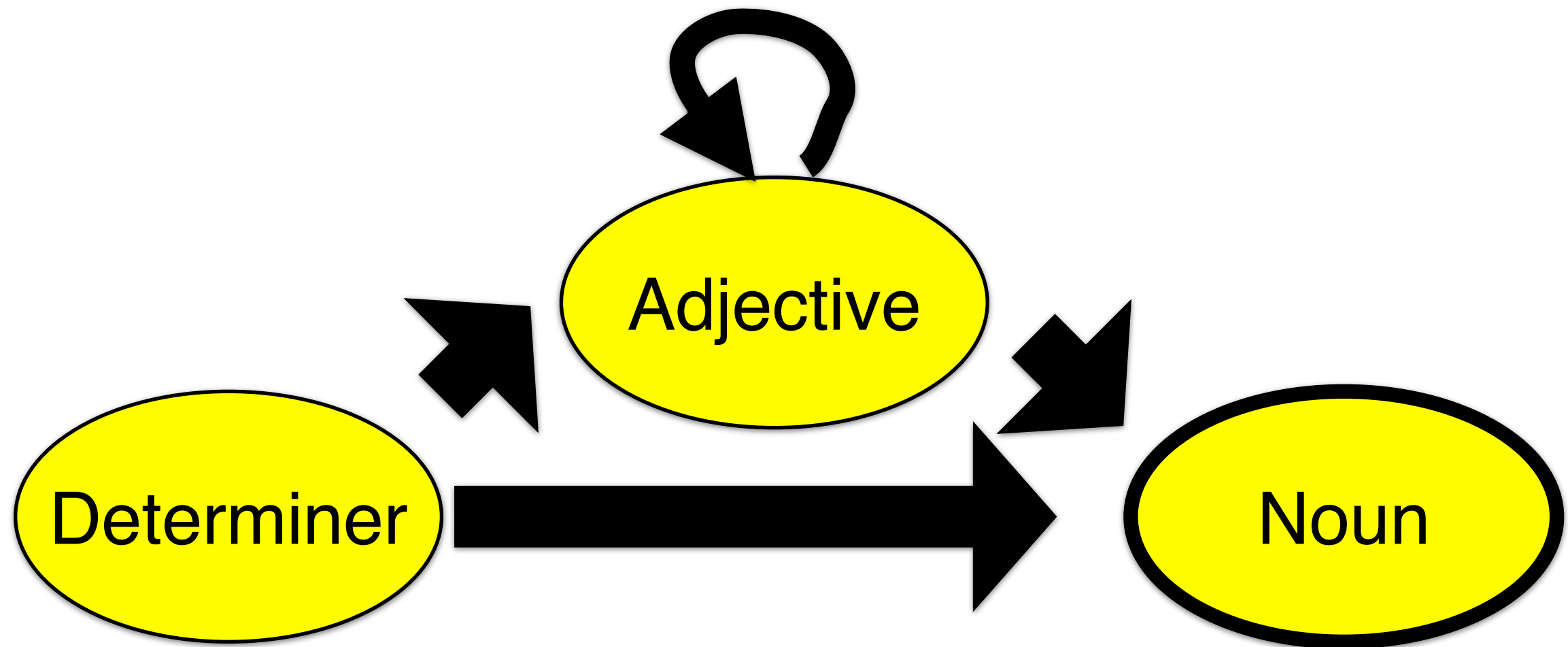
Adjectives can **modify** nouns.

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

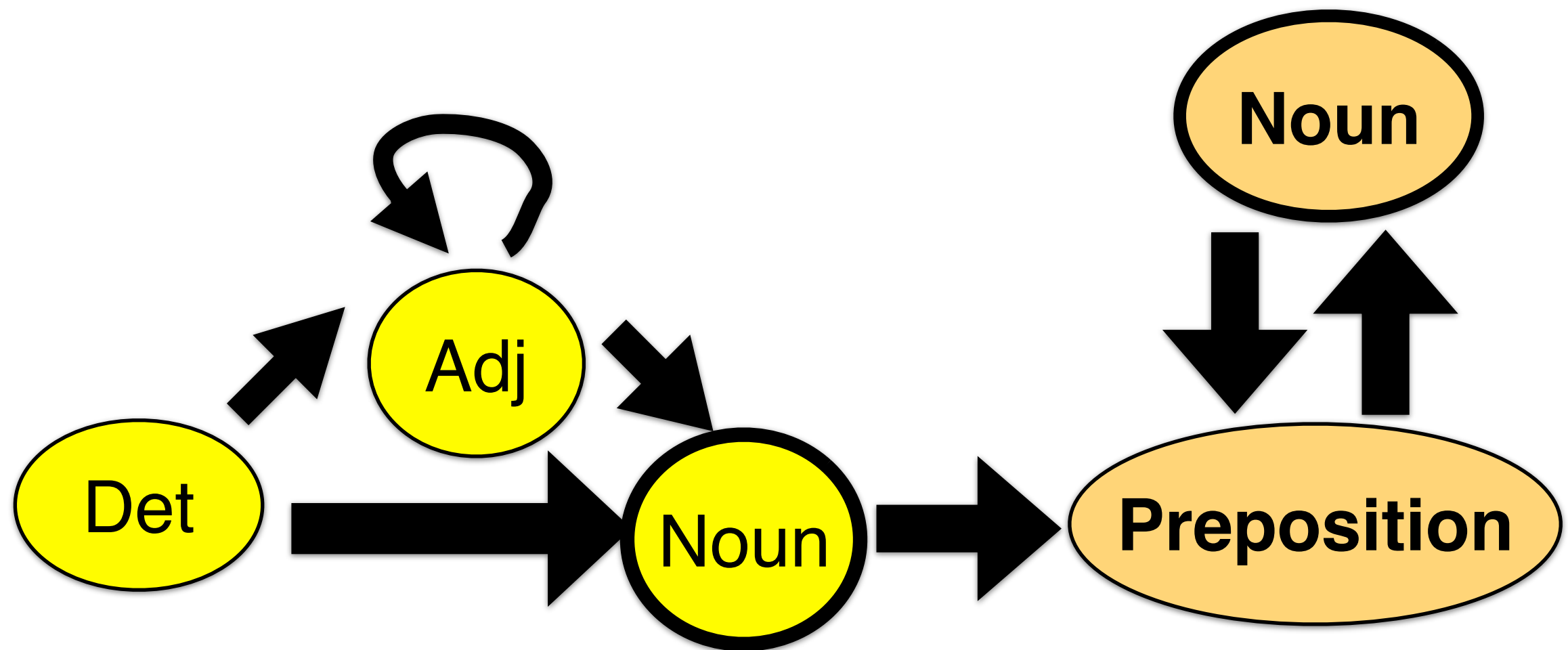
the ball in the garden

the ball in the garden behind the house

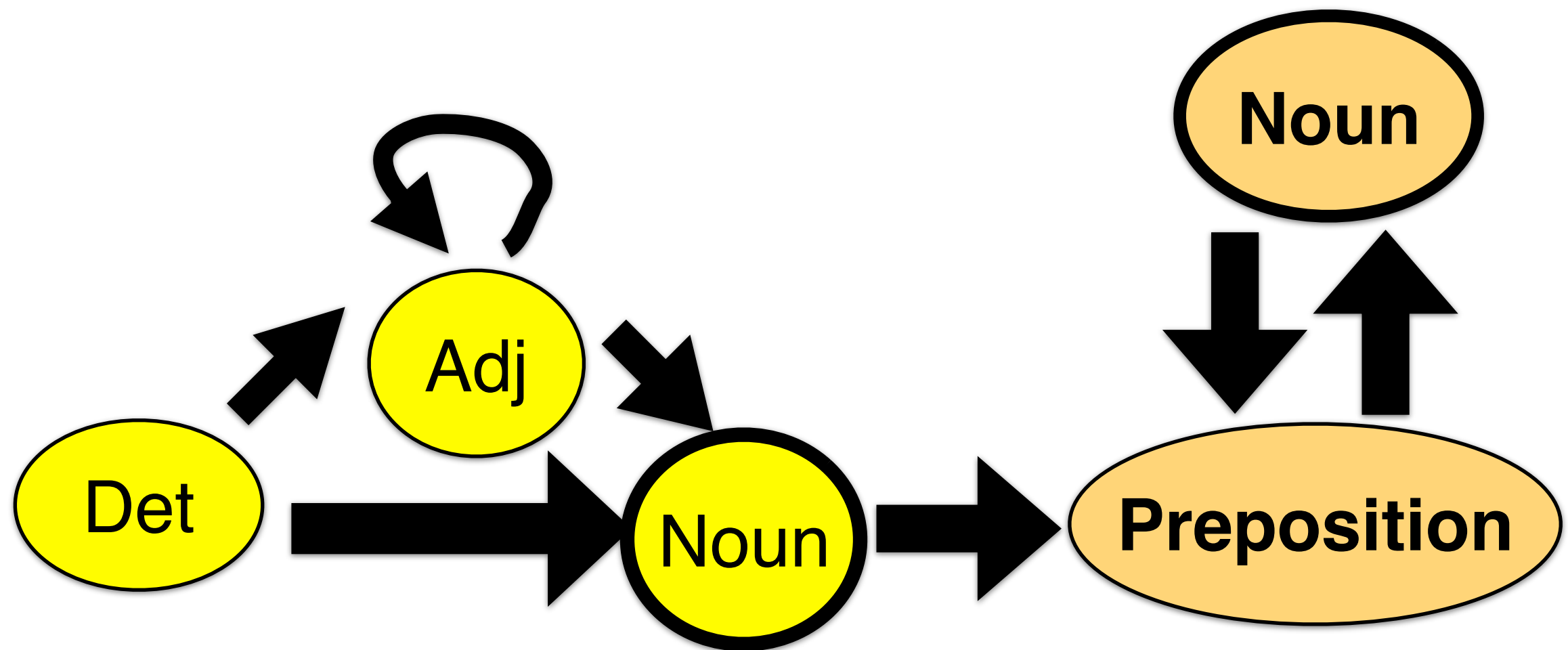
**the ball in the garden behind the house next
to the school**

....

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

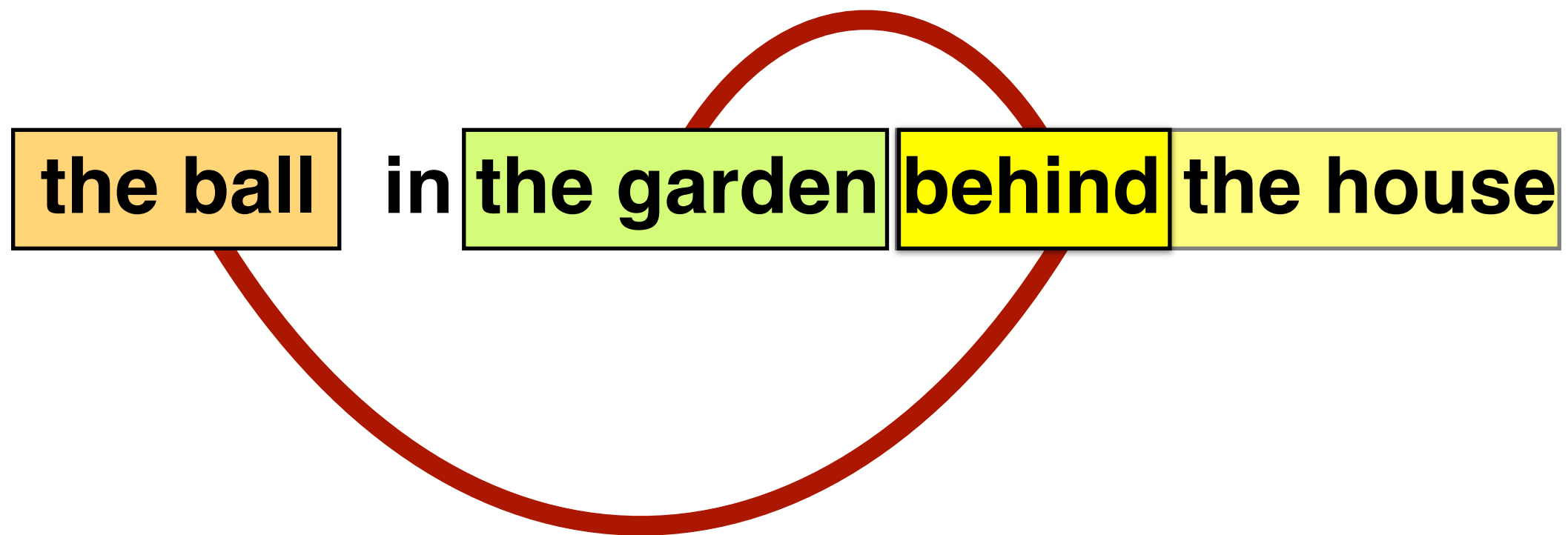
in the garden

behind

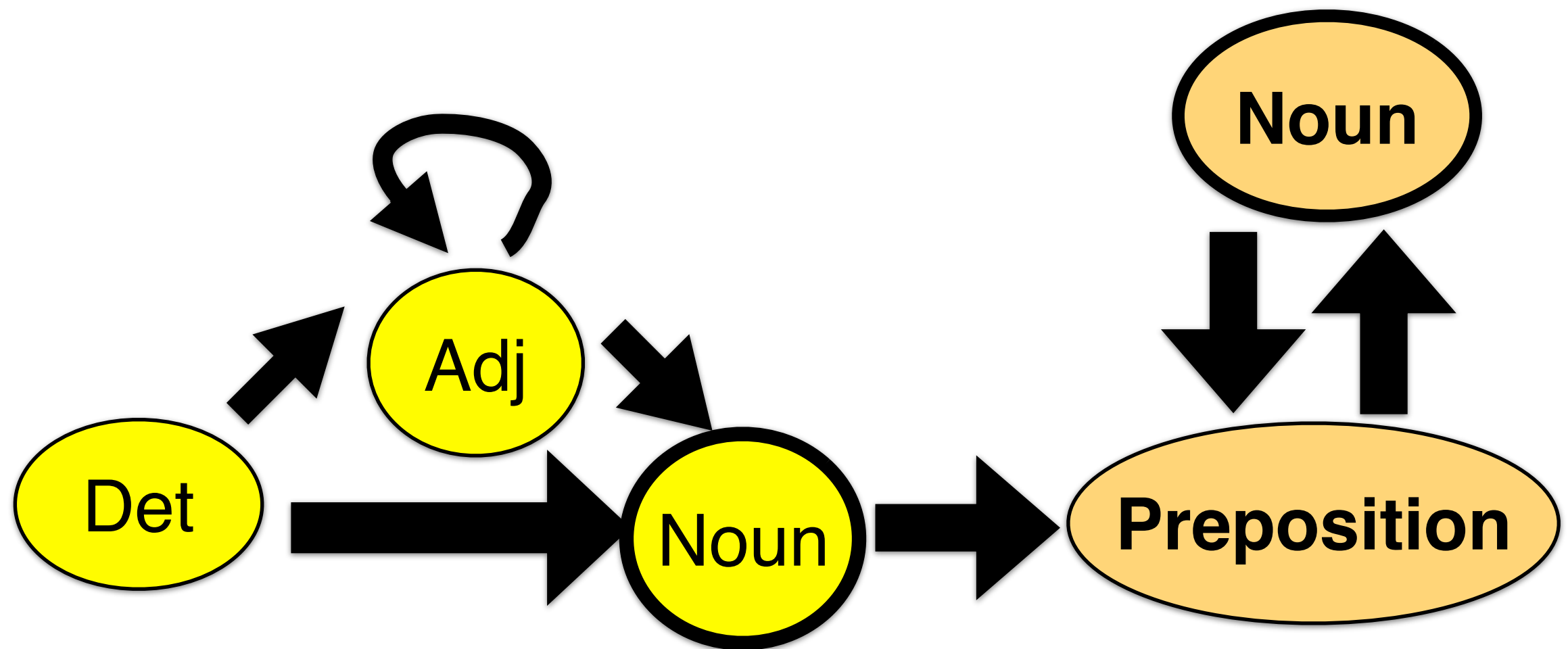
the house



What does this *mean*?



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

CFGs define parse trees

N $\rightarrow \{sushi, tuna\}$

P $\rightarrow \{with\}$

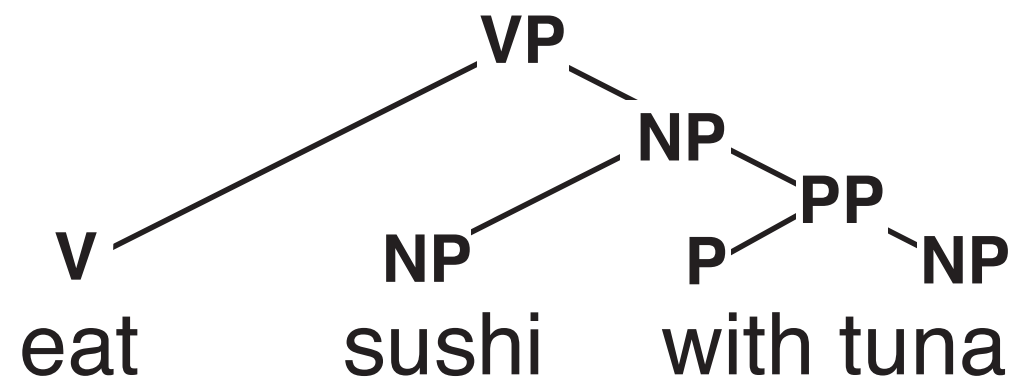
V $\rightarrow \{eat\}$

NP \rightarrow **N**

NP \rightarrow **NP PP**

PP \rightarrow **P NP**

VP \rightarrow **V NP**



Structural ambiguity

results in multiple parse trees

N → {*sushi, tuna*}

P → {*with*}

V → {*eat*}

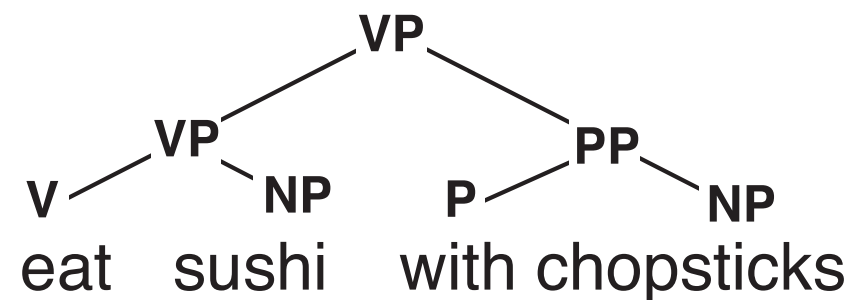
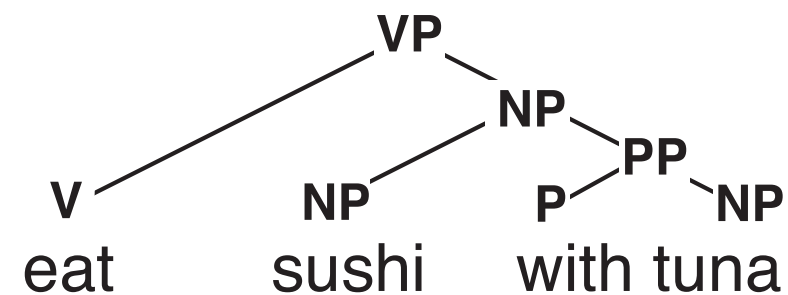
NP → **N**

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*}

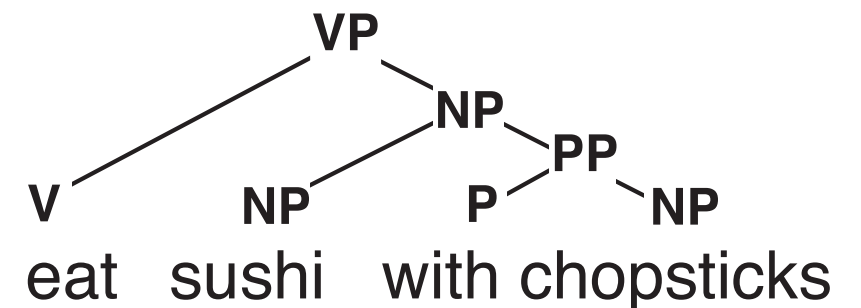
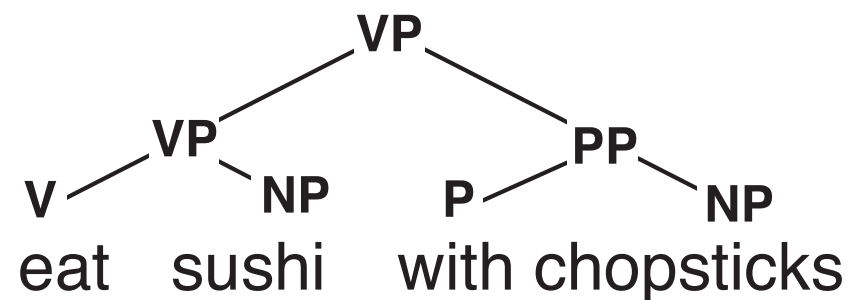
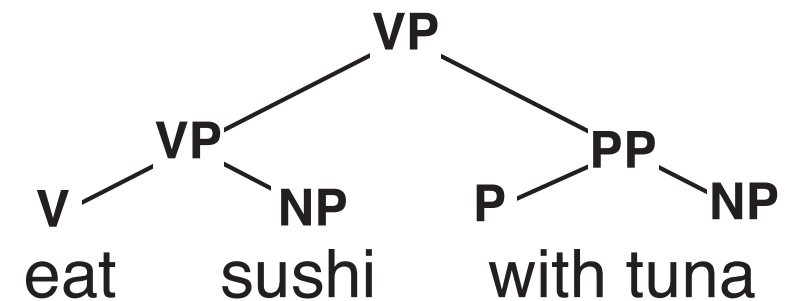
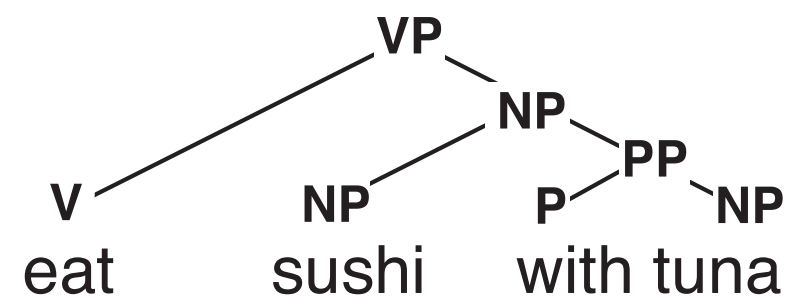
NP → **N**

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*}

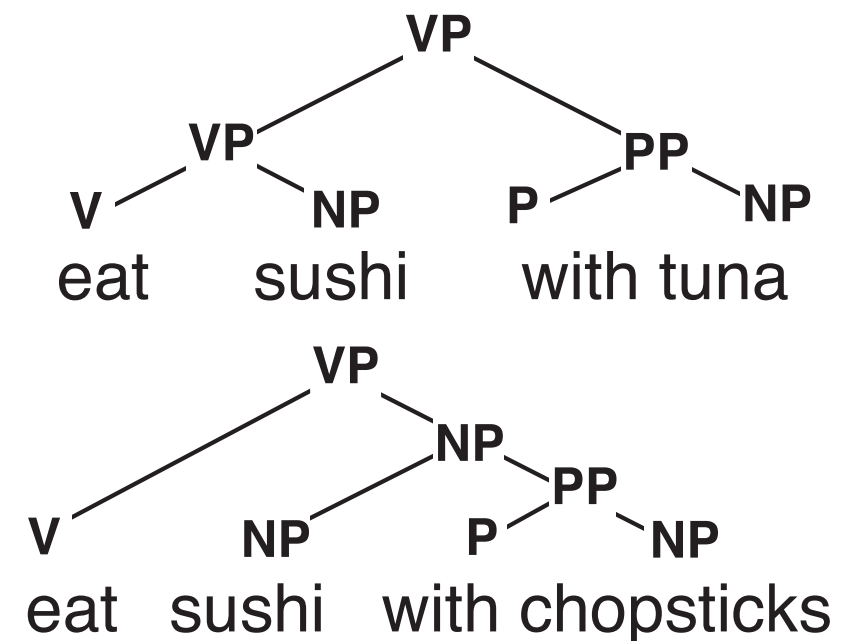
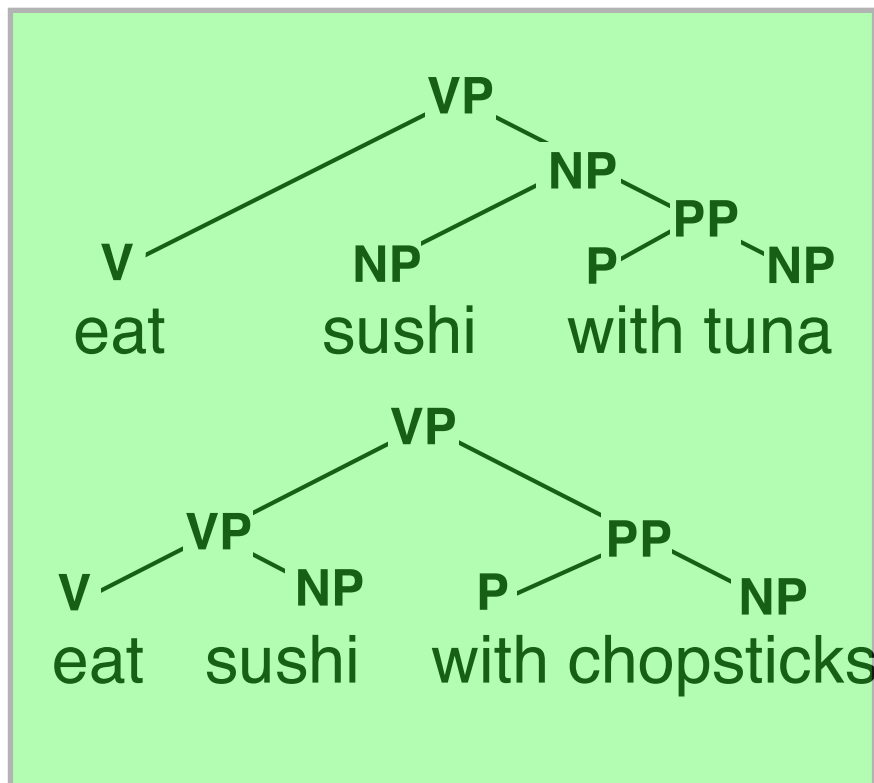
NP → **N**

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*}

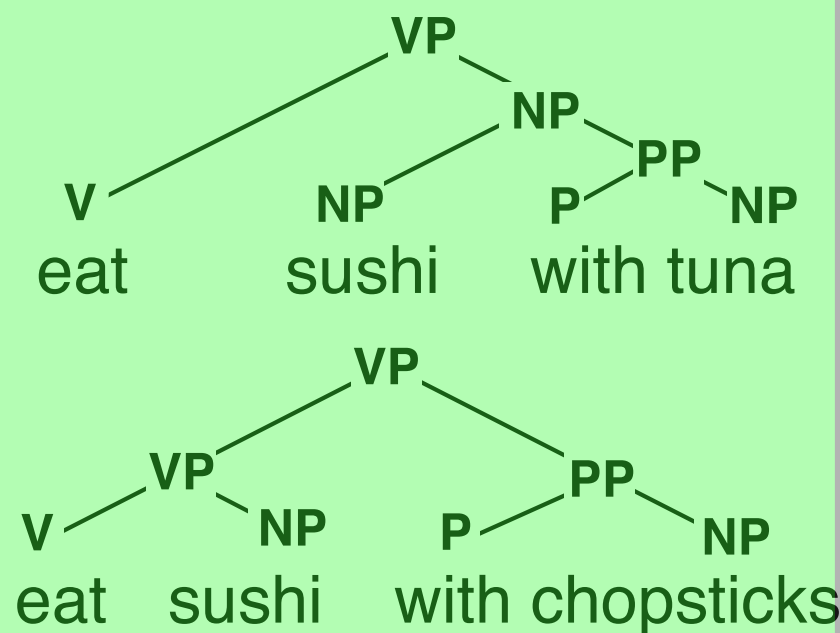
NP → **N**

NP → **NP PP**

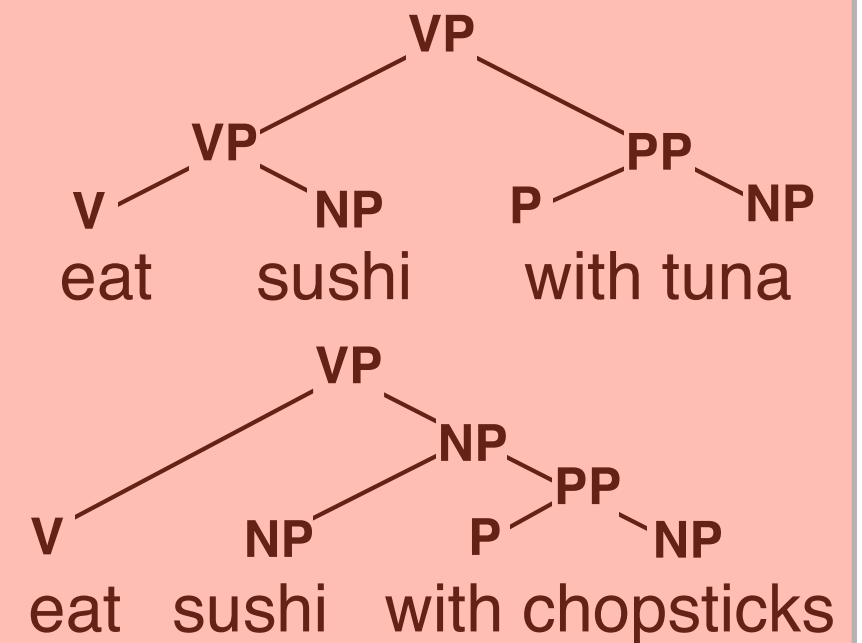
PP → **P NP**

VP → **V NP**

VP → **VP PP**



**Correct
Structures**



**Incorrect
Structures**