15-780: Graduate AI
Natural Language Processing

Geoff Gordon
with thanks to Noah Smith, LTI, MLD

Admin

- *Apologies for the late start to Tuesday's lecture!*
- *HW3 due today*
- *HW4 out (due Tuesday, 11/13)*

Admin

- *Project proposals due Thursday, 11/8*
- *Work in groups of 2*
- *If you're having trouble finding a partner, email thlin@cs by tomorrow (Friday)*
  - *include brief statement of interests*
- *Anyone who emails will get the list of people looking for partners*

Project proposals

- *A good proposal answers:*
  - *What result do you hope to get?*
  - *Why is it interesting?*
  - *Why is it related to Grad AI?*
- *Limit 1 page*

Admin

- *Midterm approaching fast!*
- *In class, Thursday 11/15 (two weeks)*
- *Review sessions: 11/12 and 11/13 in evening (time and place TBA)*

Admin

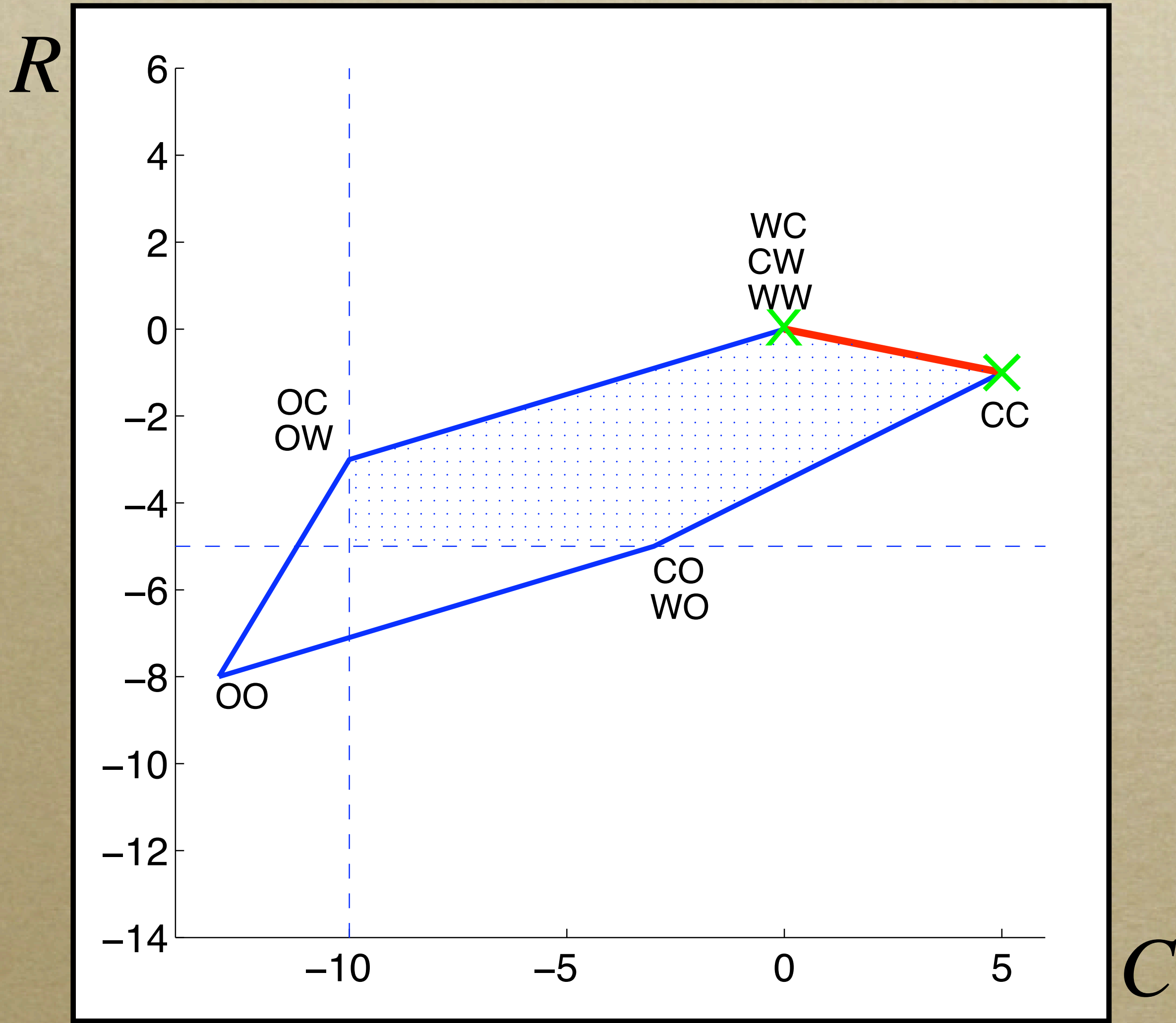
- *By request, we are also adding a hands-on practice session*
  - *scheduled for next Monday evening, time and place TBA over email*
- *Idea: work through some larger example problems in detail*
- *Not necessarily midterm-related*
- *Email me requests for problem types*

Game example

A political game

	C	W	O
C	$-1, 5$	$0, 0$	$-5, -3$
W	$0, 0$	$0, 0$	$-5, -3$
O	$-3, -10$	$-3, -10$	$-8, -13$

A political game

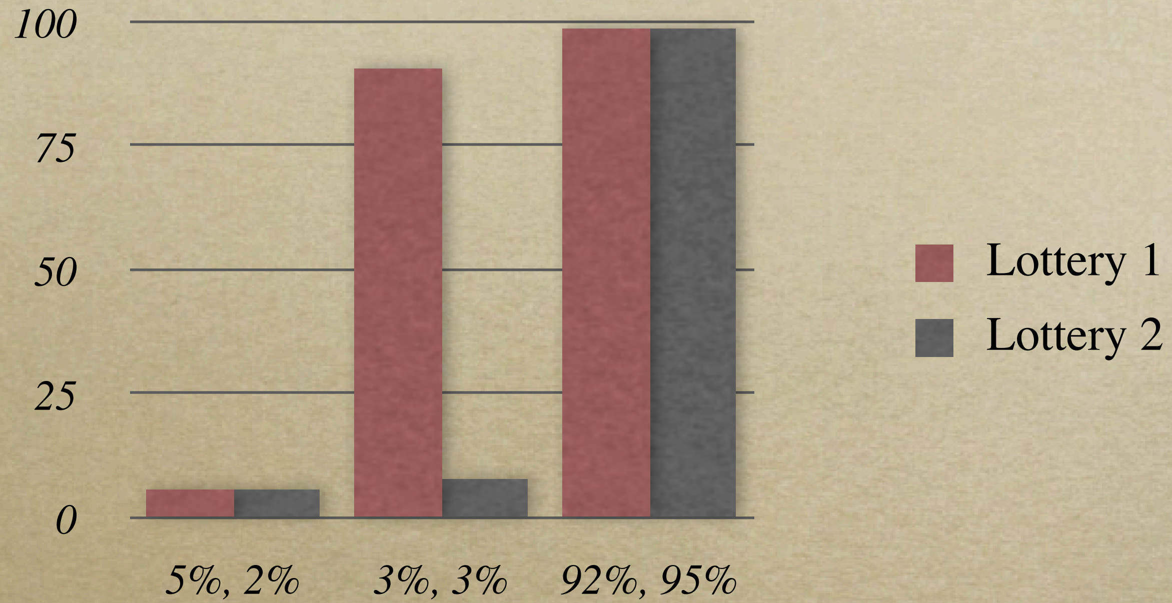


What if?

- *Didn't know each other's exact payoffs?*
- *Couldn't observe each other's exact actions?*
- *Actions altered state of world?*
- *We'll talk about some of these in later part of course*

Another example

Let’s play the lottery



- *($6, .05; $91, .03; $99, .92)*
- *($6, .02; $8, .03; $99, .95)*
- *Which would you pick?*

Rationality

- *People often pick*
  - *($6, .05; $91, .03; $99, .92)*
- *over*
  - *($6, .02; $8, .03; $99, .95)*
- *But, note* ***stochastic dominance***

Stochastic dominance

Birnbaum & Navarrete. Testing Descriptive Utility Theories: Violations of Stochastic Dominance and Cumulative Independence

NLP

(thanks to Noah Smith)

(errors are my own)

Overview

- *Overview of trends and tradeoffs in NLP*
- *Major issues in language processing*
- *Discussion of example applications, problems & solutions*
  - *Statistical parsing*
  - *Machine translation*

Language is central to intelligence

- *One of the best ways to communicate with those pesky humans*
- *One of the best ways to represent complex, imperfectly-defined concepts*
- *One of the most flexible reasoning systems ever invented*

Language is central to intelligence

Language shapes the way we think, and determines what we can think about.

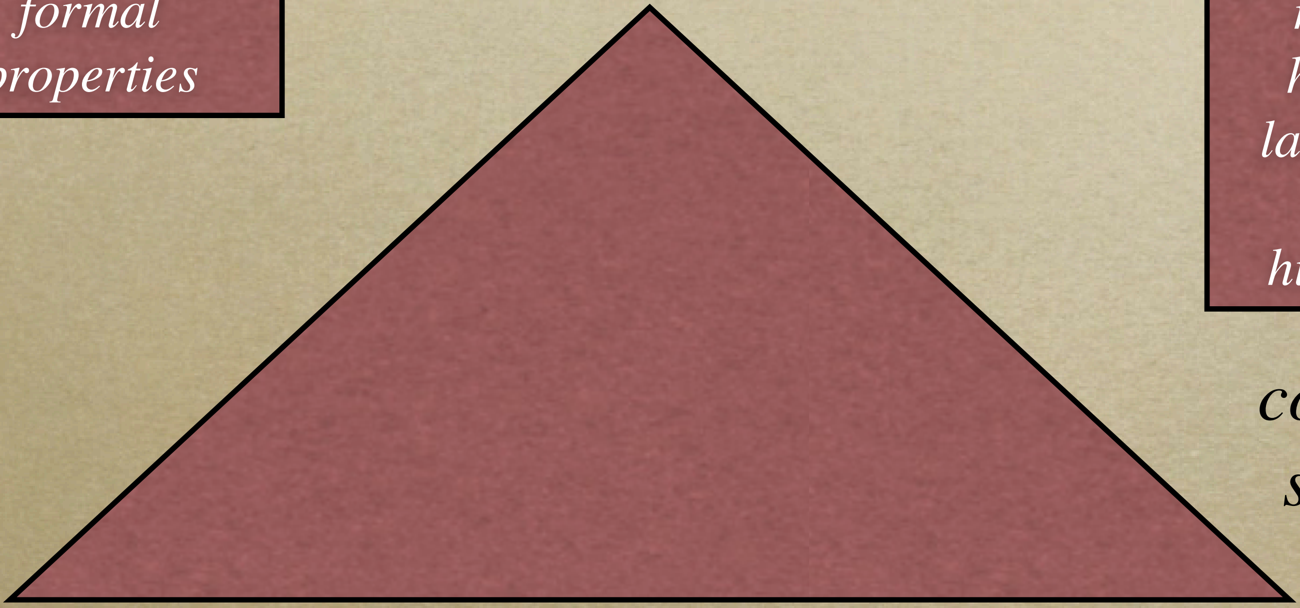
— Benjamin Lee Whorf

NLP is Interdisciplinary

goal: understand formal properties

mathematics/machine learning/CS theory

goal: plausibly model human language (like humans)



linguistics

cognitive science

goal: perfectly model human language

engineering

goal: build stuff that works

NLP is hard!

- *Ambiguity at all levels*
- *Many different styles of language*
- *Language is productive*

Levels of Language

(speech)	*(text)*
	collections
conversations	*documents*
utterances	*sentences*
words	*words*
sounds	*characters*

Boundaries not always unambiguous!

Where are the words?

世界人权宣言

联合国大会一九四八年十二月十日第*217A(III)*号决议通过并颁布

1948 年 *12* 月 *10* 日，联合国大会通过并颁布《世界人权宣言》。这一具有历史意义的《宣言》颁布后，大会要求所有会员国广为宣传，并且"不分国家或领土的政治地位，主要在各级学校和其他教育机构加以传播、展示、阅读和阐述。"《宣言》全文如下：

序言

鉴于对人类家庭所有成员的固有尊严及其平等的和不移的权利的承认，乃是世界自由、正义与和平的基础，

鉴于对人权的无视和侮蔑已发展为野蛮暴行，这些暴行玷污了人类的良心，而一个人人享有言论和信仰自由并免予恐惧

Where are the **morphemes**?

İnsan hakları evrensel beyannamesi

Önsöz

İnsanlık ailesinin bütün üyelerinde bulunan haysiyetin ve bunların eşit ve devir kabul etmez haklarının tanınması hususunun, hürriyetin, adaletin ve dünya barışının temeli olmasına,

İnsan haklarının tanınmaması ve hor görülmesinin insanlık vicdanını isyana sevkeden vahşiliklere sebep olmuş bulunmasına, dehşetten ve yoksulluktan kurtulmuş insanların, içinde söz ve inanma hürriyetlerine sahip olacakları bir dünyanın kurulması en yüksek amaçları oralak ilan edilmiş bulunmasına,

İnsanin zulüm ve baskıya karşı son çare olarak ayaklanmaya mecbur kalmaması için insan haklarının bir hukuk rejimi ile korunmasının esaslı bir zaruret olmasına,

Uluslararasında dostça ilişkiler geliştirilmesini teşvik etmenin esaslı bir zaruret olmasına,

Which words are these?

הכרזה לכל באי עולם בדבר זכויות האדם

הואיל והכרה בכבוד הטבעי אשר לכל בני משפחת האדם ובזכויותיהם
הצדק והשלום בעולם.

הואיל והזלזול בזכויות האדם וביזוין הבשילו מעשים פראיים שפגעו קש
ייהנו כל יצורי אנוש מחירות הדיבור והאמונה ומן החירות מפחד וממחסו

הואיל והכרח חיוני הוא שזכויות האדם תהיינה מוגנות בכוח שלטונו של
להשליך את יהבו על מרידה בעריצות ובדיכוי.

הואיל והכרח חיוני הוא לקדם את התפתחותם של יחסי ידידות בין האומו

Word Sense Disambiguation

… **plant** …

… *workers at the* **plant** …

… **plant** *a garden* …

… **plant** *meltdown* …

… *graze* … **plant** …

… *house* **plant** …

… *CIA* **plant** …

… **plant** *firmly on the ground* …

pick a dictionary definition for each

Language

- *Ambiguity at all levels*
  - *haven't even gone above words yet—it gets worse!*
- *Diversity of domains*
  - *New York Times v. Wall Street Journal*
  - *newspaper v. research paper*
  - *newspaper v. blog*
  - *blog v. conversation, chat, …*

Language is productive

- *New words appear all the time*
- *Very many rare words, so it's a common occurrence to see a rare word*

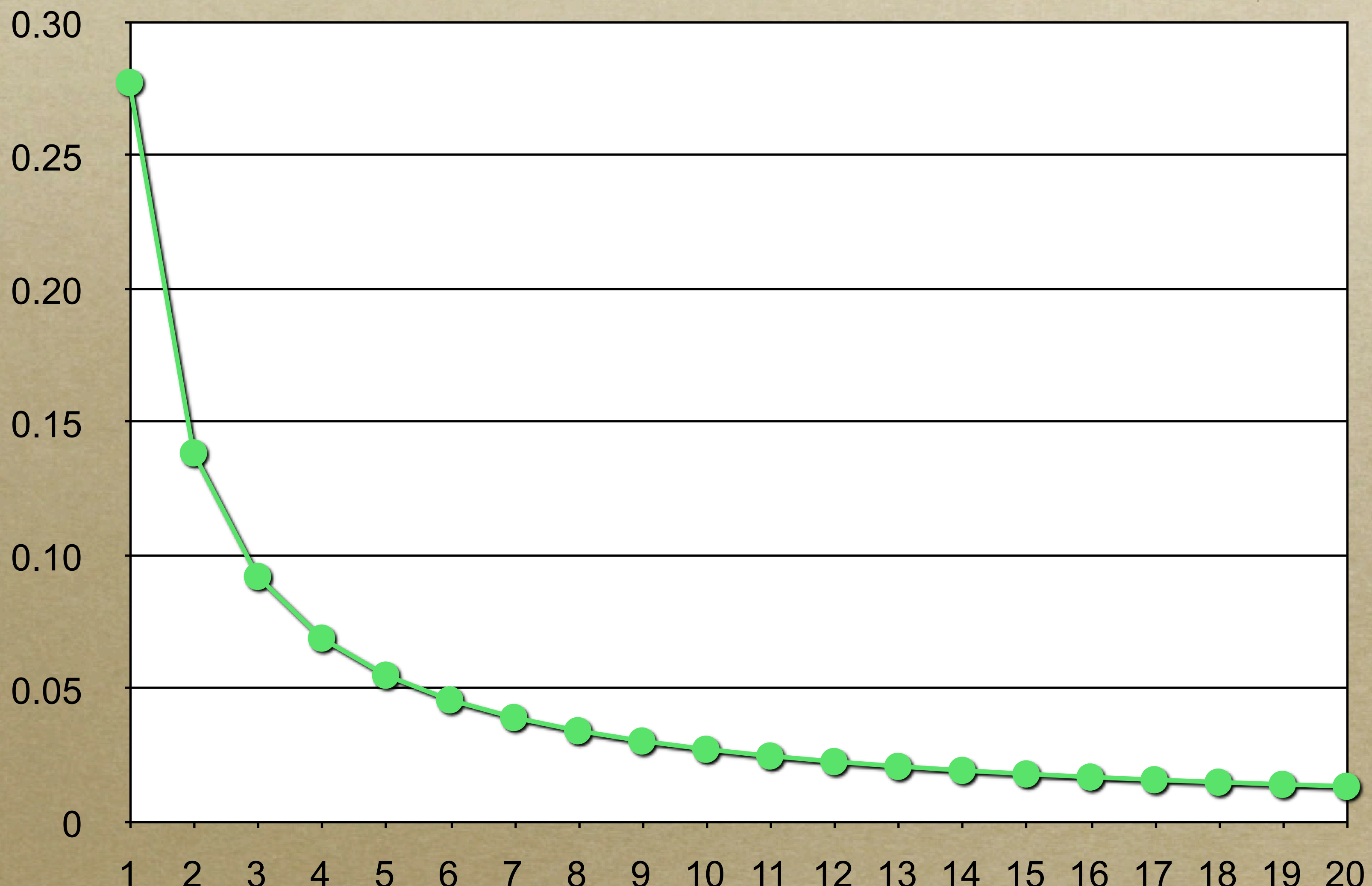
Zipf's Law

- *Type: element of a set (e.g., vocabulary word)*
- *Zipf: The most frequent types are* **extremely** *frequent, and there's a long tail.*

$$\textit{frequency} \times \textit{rank} \approx \textit{constant}$$

- *First noticed for words; holds for pretty much everything in NLP.*
- *Result:* **sparse data**, *generalization hard*

Zipf's Law


0.30
0.25
0.20
0.15
0.10
0.05
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Word Classes

- *Useful to* **abstract** *from the words themselves.*
- *Nouns: { cat, dog, horse, pig, cookie, protest, … }*
- *Verbs: { hunt, eats, kill, cook, animated, … }*
- *More verbs: { dog, horse, pig, protest, … }*
- *More nouns: { hunt, eats, kill, cook, … }*
- *Adjectives: { animate, funny, heavy-handed, … }*
- *Linguist required: { what, that, up, umm, … }*

Word Classes

- *Haven't even gotten to fancier classes:*
- *Animate: { horse, cat, professor, … }*
- *Place: { New York, Wean 5409, under the boardwalk, … }*
- *Intangible: { blue, filibuster, complexity, … }*

Goals

- *Given all of this complexity and ambiguity, we want to:*
    - *Understand*
    - *Respond*
    - *Translate*
    - *Classify*
    - …

Goals

For any of these goals, we need some **deeper** *representation.*

Two common levels beyond words:

Words ➔ Syntax ➔ Semantics

How do we get there?

Syntax

- *First thought: use* `lex/yacc` *to build a* **parser** *for a natural language just like for a programming language!*
  - *Need to* **know** *grammar of NL*
  - *Tremendous number of possible rules; no spec.*
    - *Zipf's law attacks again.*
- *Where is NL in the Chomsky Hierarchy?*

Chomsky Hierarchy

Grammar type	*Machine type*
Unrestricted	*Turing machine*
Context-sensitive	*Nondeterministic linear-bounded automaton*
Context-free	*Pushdown automaton*
Regular	*Finite-state machine*

Chomsky Hierarchy

- *Some linguistic phenomena don't exhibit very long-ranging influences.*
  - *Phonetics, phonology.*
  - *Speech recognition uses mostly* **finite-state** *models.*
- *Linguists have used examples to demonstrate that there is*
  - *arbitrary center-embedding (i.e., NL is not FS)*

NL is not Finite-State

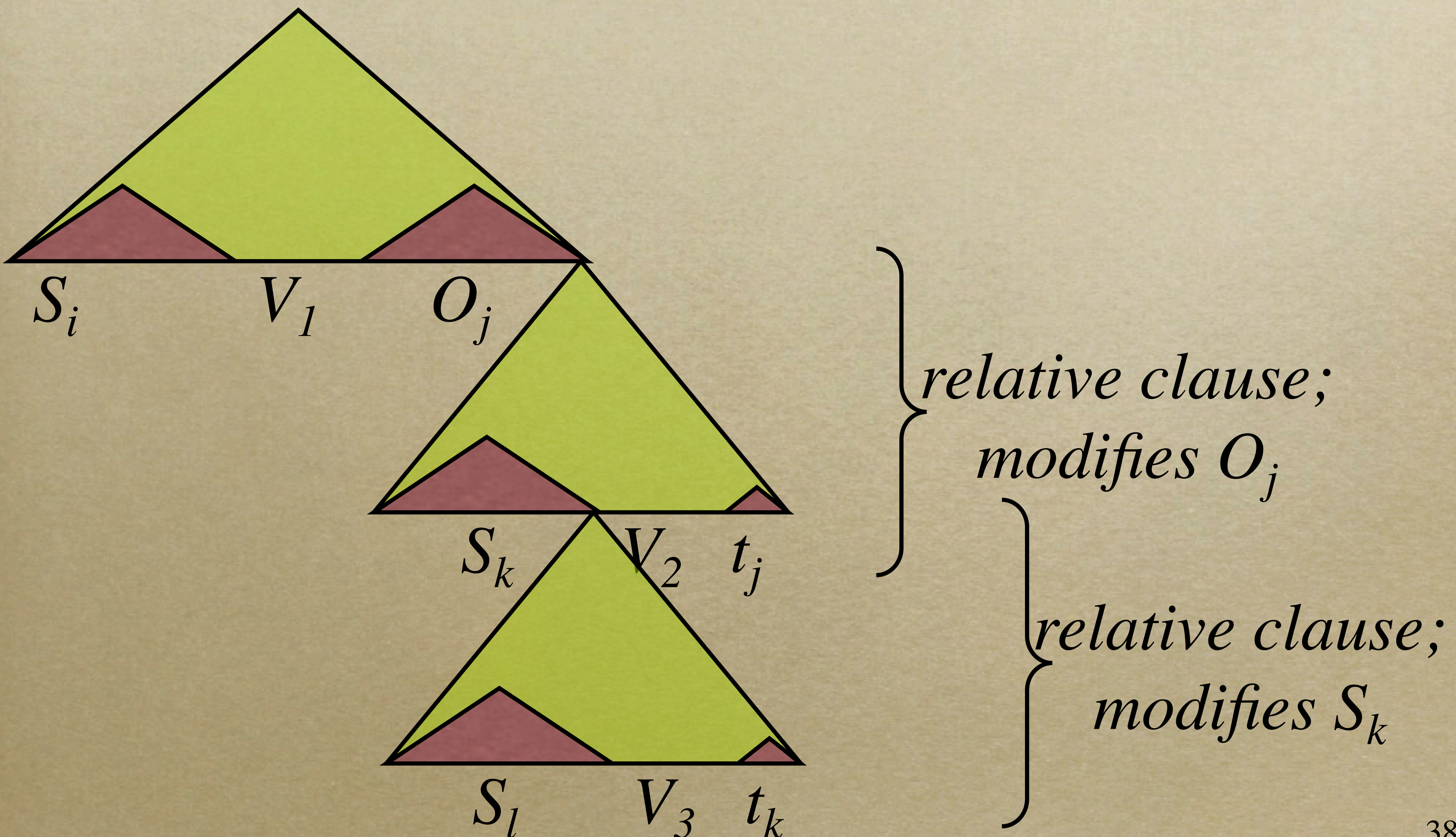
This is the cat.

This is the dog the cat chased.

This is the man the dog the cat chased bit.

This is the woman the man the dog the cat chased bit kissed.

NL is not Finite-State



Chomsky Hierarchy

- *Some linguistic phenomena don't exhibit very long-ranging influences.*
  - *Phonetics, phonology.*
  - *Speech recognition uses mostly* **finite-state** *models.*
- *Linguists have used examples to demonstrate that there is*
  - *arbitrary center-embedding (i.e., NL is not FS)*
  - *cross-serial dependencies (i.e., NL is not CF).*

Chomsky Hierarchy

- *Many context-sensitive models of language have been proposed!*
- *NLP still uses FS or CF models mostly, for speed and coverage.*

Syntax

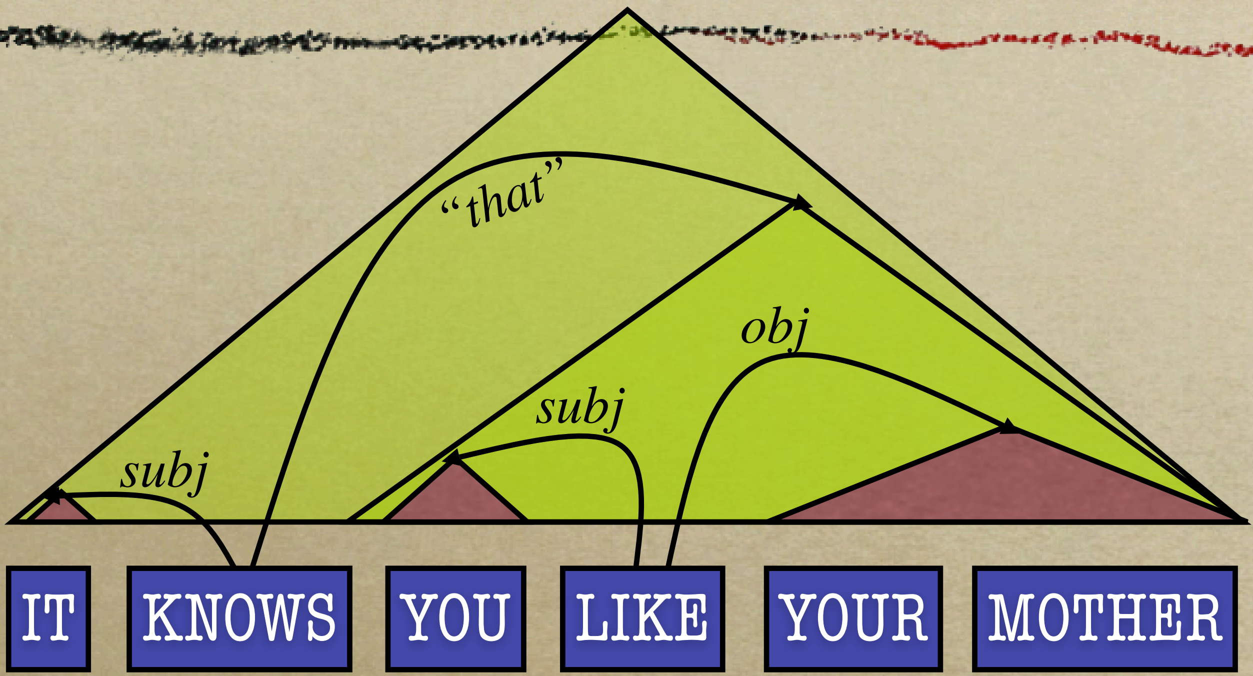
- *First thought: use* `lex/yacc` *to build a* **parser** *for a natural language just like for a programming language!*
- *Where is NL in the Chomsky Hierarchy?*
  - *Context-sensitive, but we'll pretend context-free.*
- *Problem: ambiguity.*

Ambiguity in English

- IT KNOWS YOU LIKE YOUR MOTHER
- IRAQI HEAD SEEKS ARMS
- JUVENILE COURT TO TRY SHOOTING DEFENDANT
- KIDS MAKE NUTRITIOUS SNACKS
- BRITISH LEFT WAFFLES ON FALKLAND ISLANDS
- LITTLE HOPE GIVEN BRAIN-DAMAGED CHILD
- NEVER WITHHOLD HERPES INFECTION FROM LOVED ONE
- STOLEN PAINTING FOUND BY TREE

Thanks to: J. Eisner, L. Lee

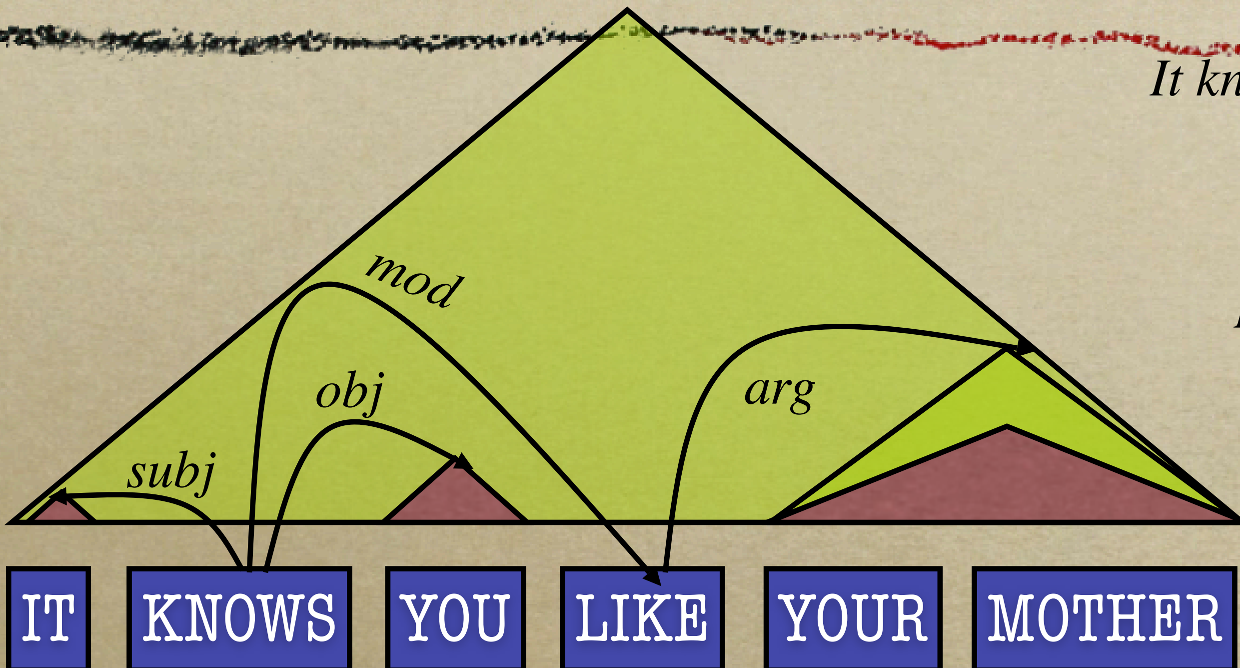
Syntactic Ambiguity



p = You like your mother.

It knows p.

Syntactic Ambiguity

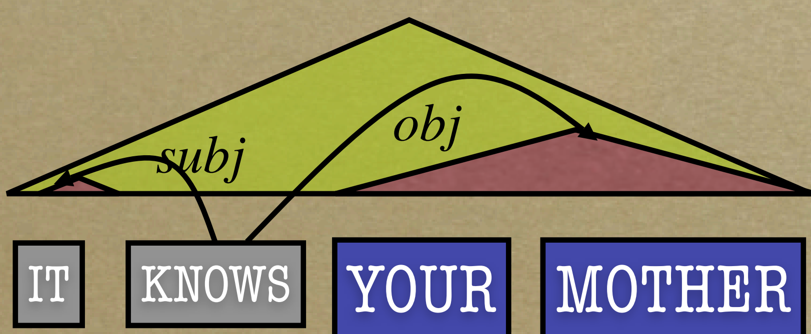


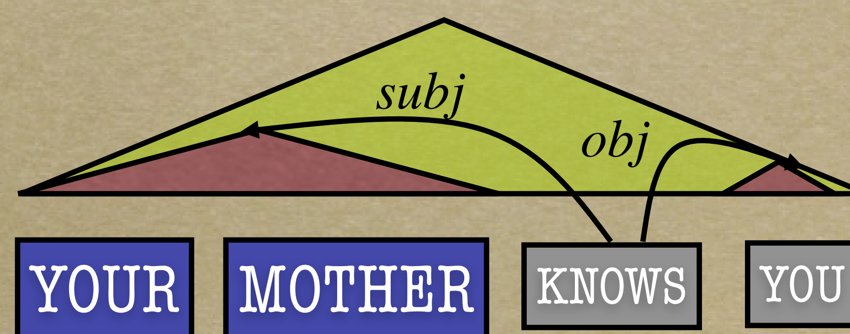
It knows your mother in a way p.

It knows you in way p.

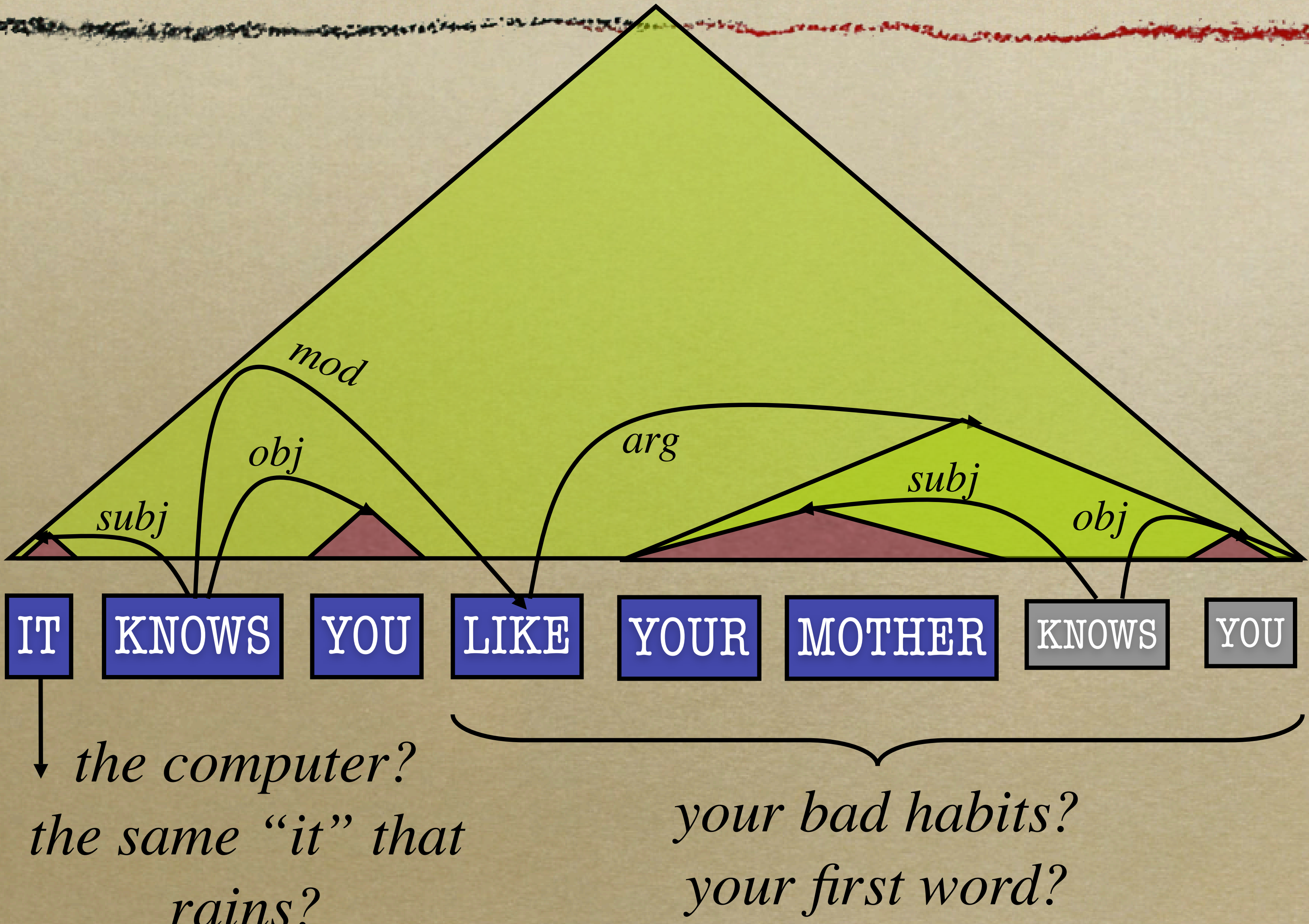
Your mother knows you in a way p.

It knows you in way p.





Semantic Ambiguity



Pragmatics and World Knowledge

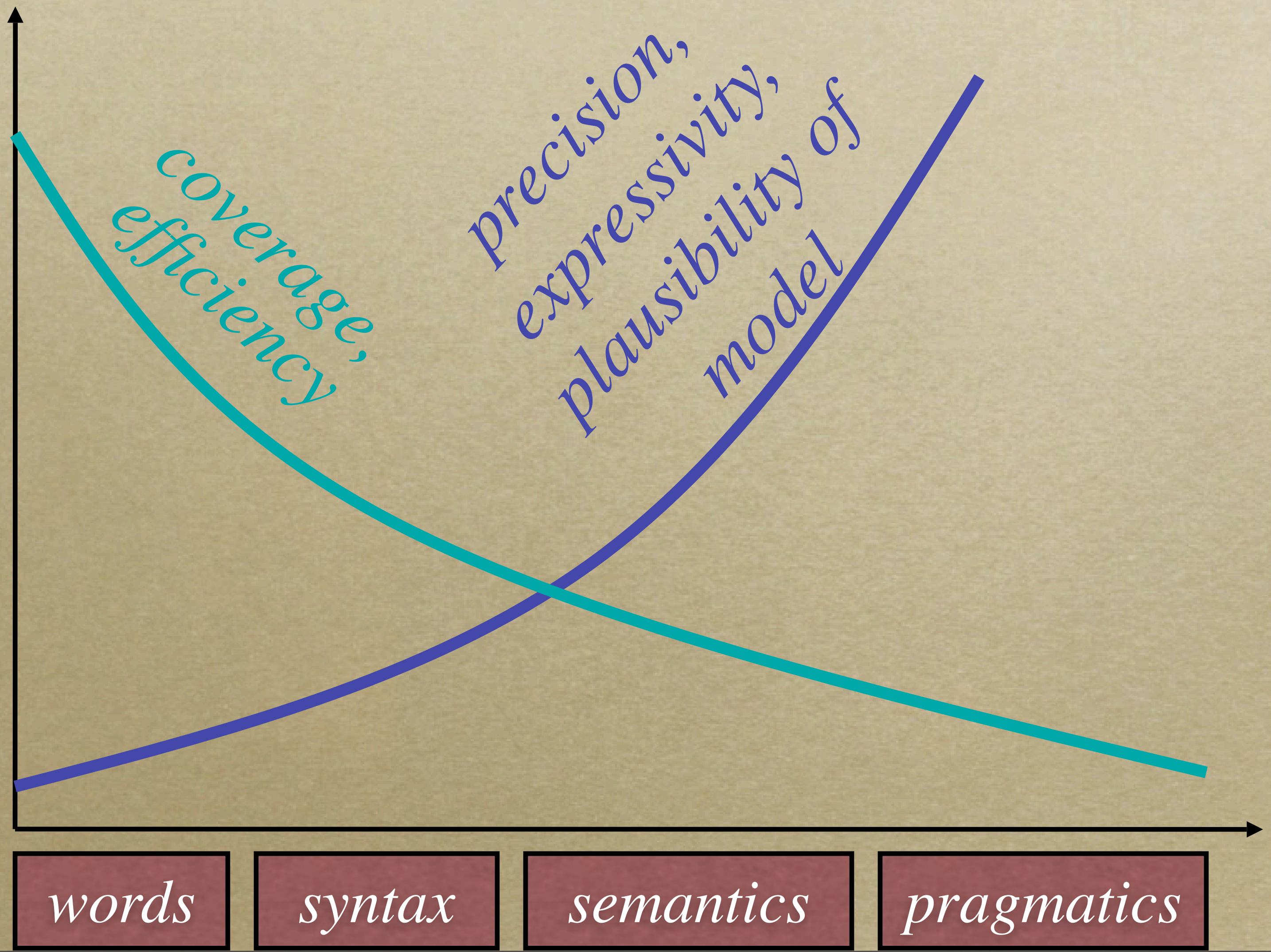
IT KNOWS YOU LIKE YOUR MOTHER

- *This statement isn't meant literally!*
- *Someone is trying to sell you something.*
- *They are juxtaposing the product with a competitor's product that is impervious to the user.*

Ambiguity

- *Headlines are more ambiguous than most text*
- *But, with any broad-coverage grammar, almost any sentence of reasonable complexity will be ambiguous, often in ways humans will never notice*

Tradeoffs


coverage,
efficiency
precision,
expressivity,
plausibility of
model
words
syntax
semantics
pragmatics

How can we handle ambiguity?

- *Probability*

The Revolution

- *In 1980s and 1990s, NLP started borrowing from speech recognition, information theory, and machine learning.*
  - *Probability models, including* **weighted grammars**
  - *Use of statistics on data (corpora)*

The Revolution

- *The new paradigm involves learning to accomplish tasks accurately from data.*
  - *How much data? What kind?*
  - *Same tradeoffs as before, and some new ones!*

Example: Statistical Parsing

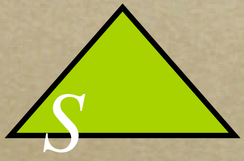
- *Input: a sentence*
- *Output: a parse tree (usually labeled constituents)*
- *Evaluation: compare to gold standard tree, count erroneous constituents.**
- *Before 1993: write rules by hand.*
- *1993: Penn Treebank*
  - *A million words' worth of* Wall Street Journal *text, annotated by linguists with a consensus structure*

How We Do It

- *Assume a model p(***t***,* **w***).*
  - *Probability distribution over discrete structures (trees and sequences).*
  - *Starting point: Probabilistic Context-Free Grammar (aka Stochastic CFG)*

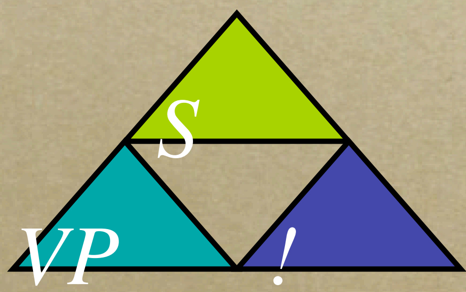
Probabilistic CFG

Just like CFGs, but with probability distribution at each rewrite.



Probabilistic CFG

Just like CFGs, but with probability distribution at each rewrite.



$p(NP\ VP\ .\ |\ S) = 0.44$

$p(VP\ !\ |\ S) = 0.26$

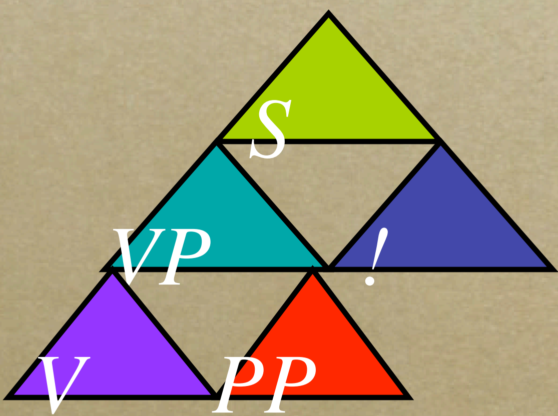
$p(Is\ NP\ VP\ ?\ |\ S) = 0.27$

$p(NP\ .\ |\ S) = 0.01$

…

Probabilistic CFG

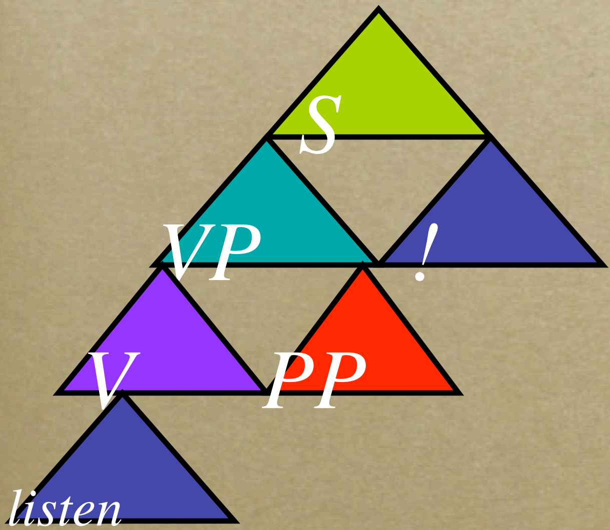
Just like CFGs, but with probability distribution at each rewrite.



$$p(V \mid VP) = 0.24$$
$$p(V\ NP \mid VP) = 0.23$$
$$p(V\ PP \mid VP) = 0.21$$
$$p(V\ NP\ PP \mid VP) = 0.16$$
$$\ldots$$

Probabilistic CFG

Just like CFGs, but with probability distribution at each rewrite.



$p(eat \mid V) = 0.03$

$p(buy \mid V) = 0.03$

$p(sell \mid V) = 0.03$

$p(implement \mid V) = 0.02$

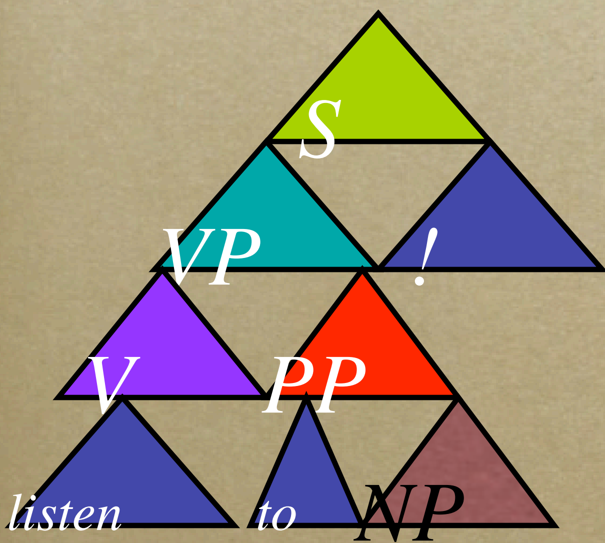
…

$p(listen \mid V) = 0.01$

…

Probabilistic CFG

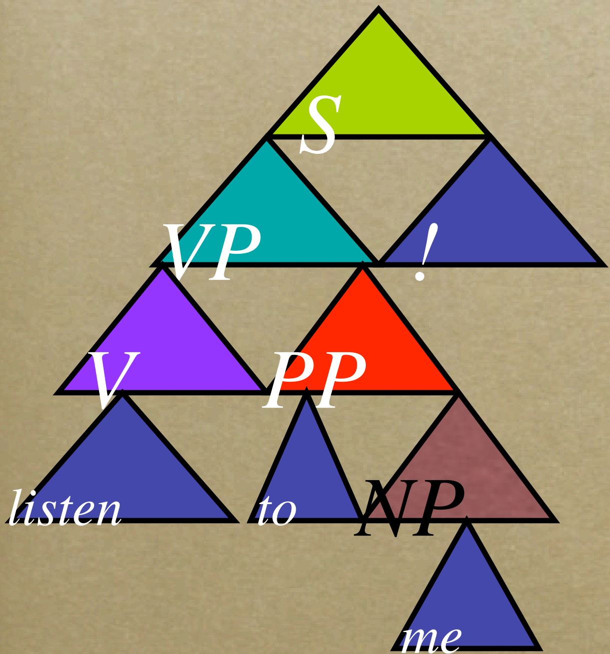
Just like CFGs, but with probability distribution at each rewrite.



$$p(\textit{to NP} \mid PP) = 0.80$$
$$p(\textit{up} \mid PP) = 0.02$$
$$\ldots$$

Probabilistic CFG

Just like CFGs, but with probability distribution at each rewrite.

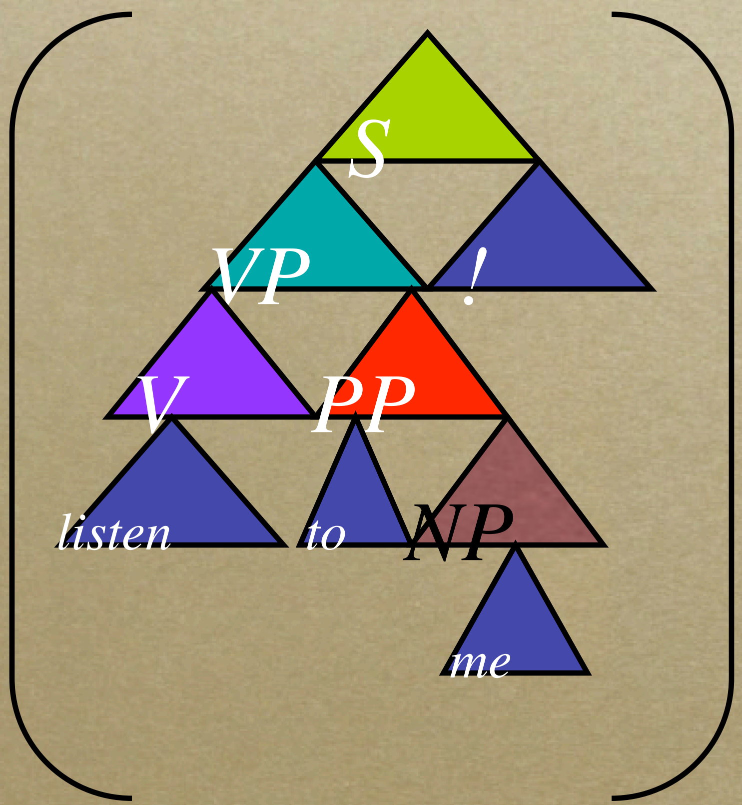


$$p(\textit{him} \mid NP) = 0.06$$
$$p(\textit{us} \mid NP) = 0.02$$
$$\ldots$$
$$p(\textit{me} \mid NP) = 0.01$$

Probabilistic CFG

Just like CFGs, but with probability distribution at each rewrite.

p



$$
\begin{aligned}
&= \quad p(VP\ !\ |\ S) && = \quad 0.26 \\
&\times\, p(V\ PP\ |\ VP) && \times\, 0.21 \\
&\times\, p(listen\ |\ V) && \times\, 0.01 \\
&\times\, p(to\ NP\ |\ PP) && \times\, 0.80 \\
&\times\, p(me\ |\ NP) && \times\, 0.01
\end{aligned}
$$

$$= 0.0000004368$$

How We Do It

- *Assume a model $p(\mathbf{t}, \mathbf{w})$.*
- *Train the model on the Treebank.*
- *To parse infer the best tree: $\max_{\mathbf{t}} p(\mathbf{t} \mid \mathbf{w})$*
    - *Discrete optimization problem*
- *Or, infer properties of posterior over trees: P(words 5–9 are a VP)*
    - *Probabilistic inference problem*

Problem

- *Problem: possible* **t** *is O(exp |***w***|)*
- *Solution: dynamic programming*
- *Similar to forward-backward or Viterbi algorithms for HMMs*
- *Analog of forward-backward:* ***inside-outside***
- *Analog of Viterbi: PCKY (Probabilistic Cocke-Kasami-Younger)—will show here*

Optimization

- *For simplicity, assume grammar in* ***Chomsky normal form***
- *All productions are*
  - $A \rightarrow BC$
  - $A \rightarrow word$
  - $A \rightarrow \varepsilon$ *(nothing)*

Chomsky normal form example

- $S \rightarrow NP\ VP$
- $VP \rightarrow V\ NP$
- $NP \rightarrow D\ N$
- $D \rightarrow$ *the* (0.6) | *those* (0.4)
- $N \rightarrow$ *cat(s)* (0.3) | *dog(s)* (0.7)
- $V \rightarrow$ *hear(s)* (0.9) | *dog(s)* (0.1)

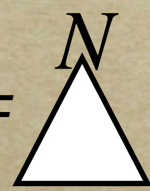
the cat dogs the dogs

Optimization

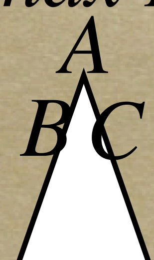
- *Given a string of nonterminals:*
  - *the cat dogs the dogs*
- *And a probabilistic context free grammar (previous slide)*
- *Figure out most likely parse*

Dynamic programming

- *String of words X*
- *For nonterminal N, write*
  - $P_{max}(N, X) = max\ P(t, X)$

    $t =$ 

- *Similarly, for production A → B C, write*
  - $P_{max}(A \rightarrow B\ C, X) = max\ P(t, X)$

    $t =$ 

Dynamic programming

- *Now, we have*
- $P_{max}(VP, X) =$

  $$\max_{VP \to \ldots} P_{max}((VP \to \ldots), X)\, P(VP \to \ldots)$$

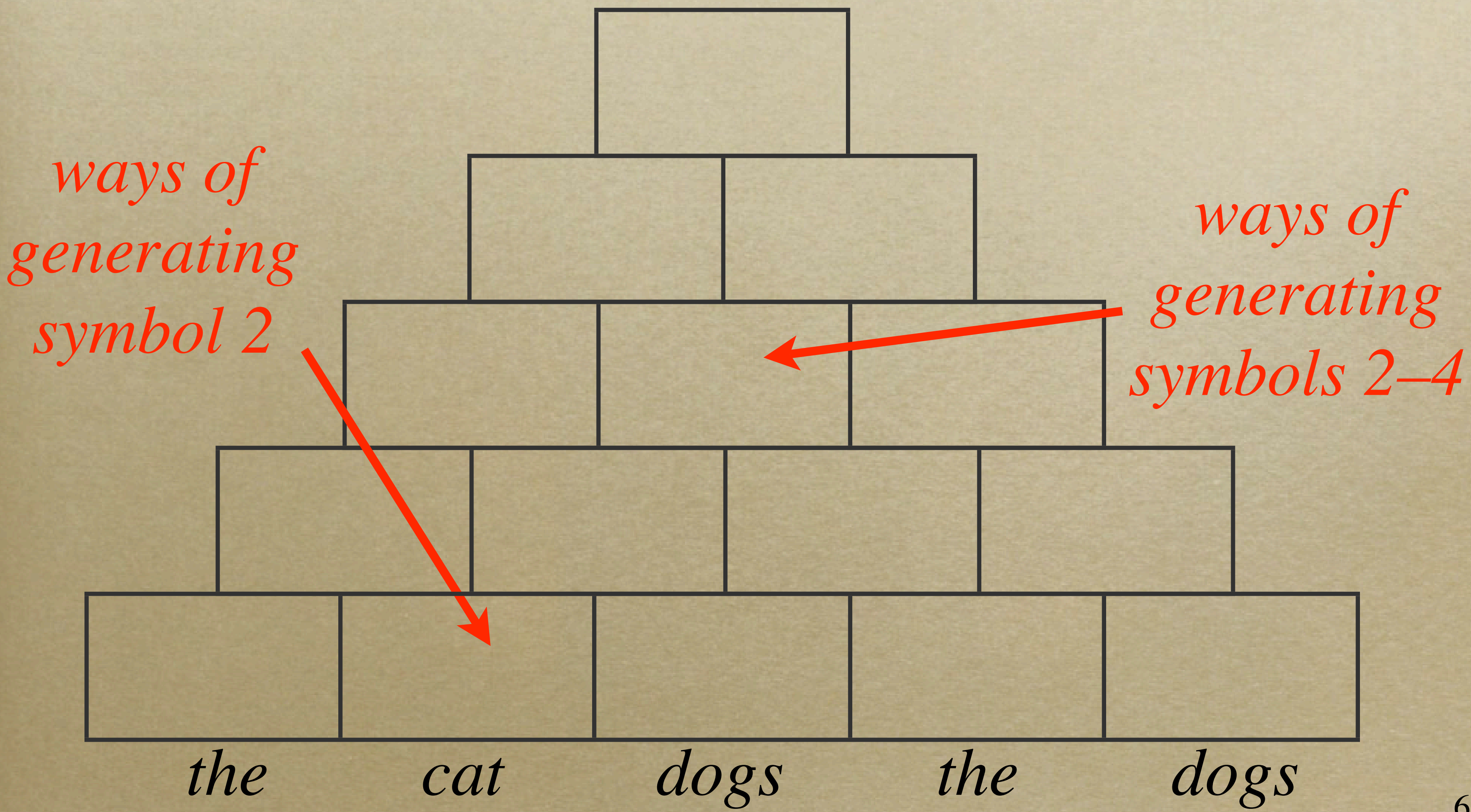
- $P_{max}((VP \to V\ NP), X) =$

  $$\max_{X=YZ} P_{max}(V, Y)\, P_{max}(NP, Z)$$

Dynamic programming

- *Build a table P(i, j, k) = probability of generating the substring from word i to word j from nonterminal k using best possible tree =* $P_{max}(k, X[i..j])$
- *In our example (5 words, 6 nonterminals), this is double[5, 5, 6]*
  - *some elements unused (triangle array)*

Dynamic programming



Dynamic programming

D → *the* (0.6) | *those* (0.3)

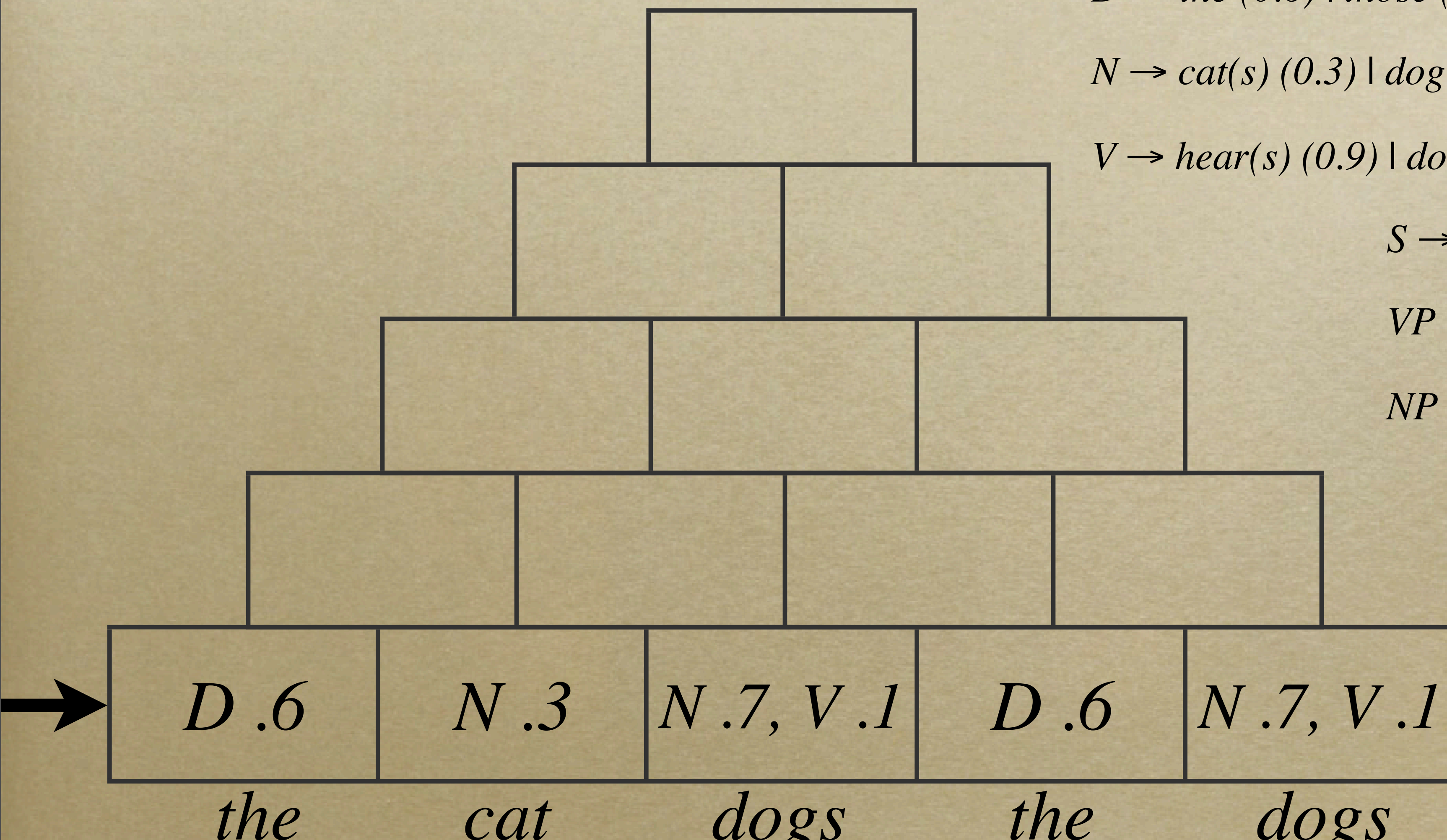
N → *cat(s)* (0.3) | *dog(s)* (0.7)

V → *hear(s)* (0.9) | *dog(s)* (0.1)

S → *NP VP*

VP → *V NP*

NP → *D N*



Dynamic programming

$D \rightarrow$ *the* (0.6) | *those* (0.3)

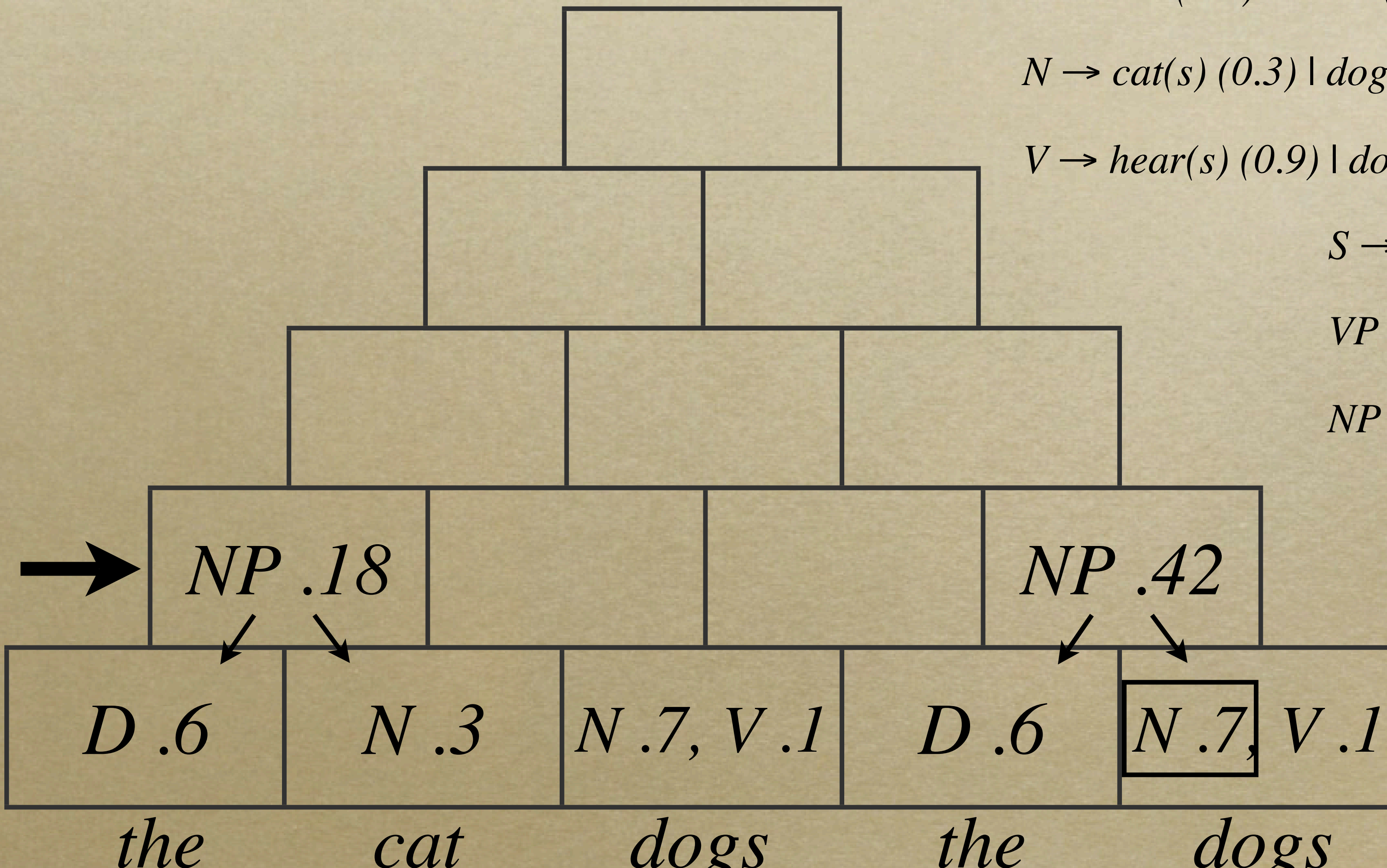
$N \rightarrow$ *cat(s)* (0.3) | *dog(s)* (0.7)

$V \rightarrow$ *hear(s)* (0.9) | *dog(s)* (0.1)

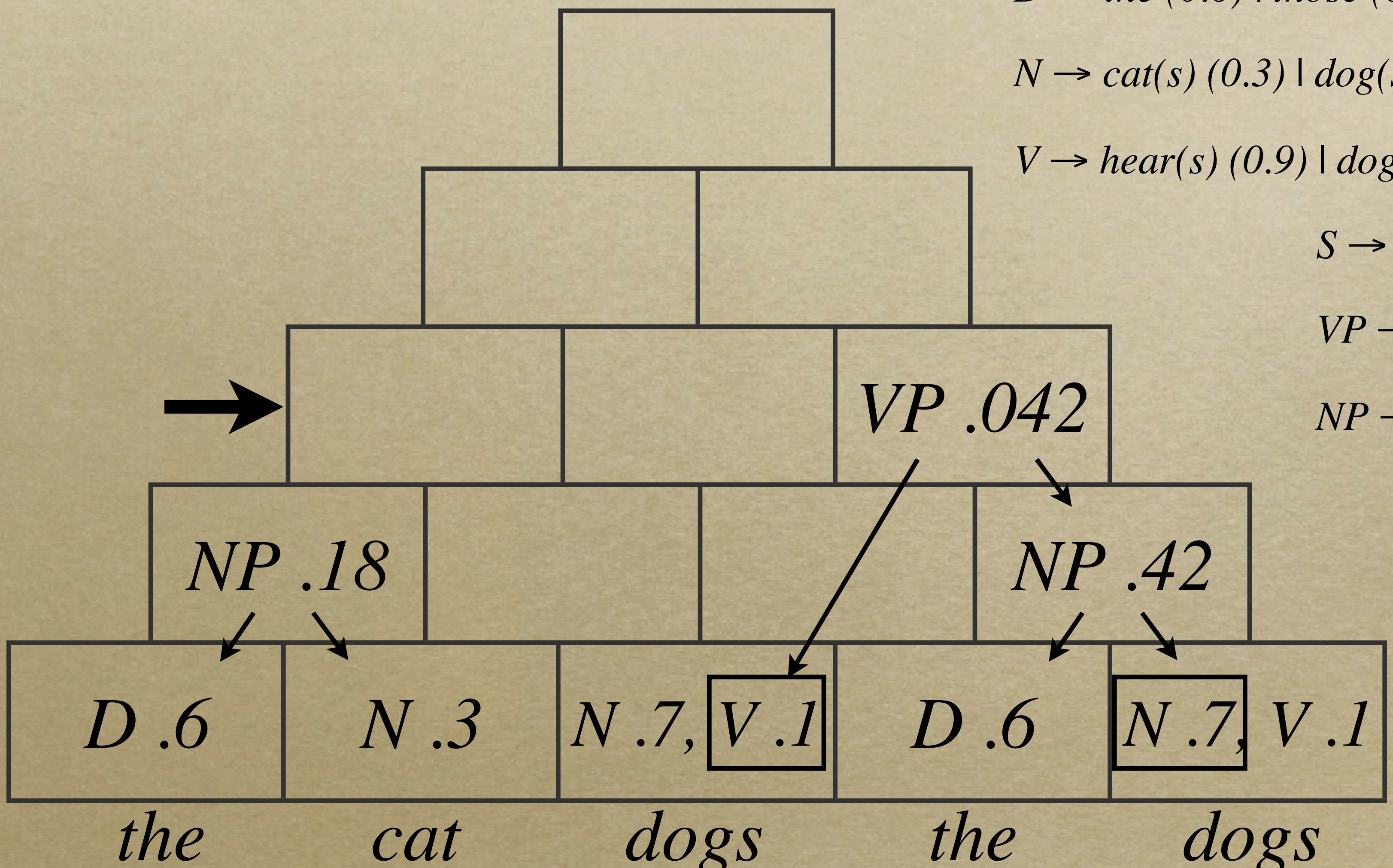
$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow D\ N$



Dynamic programming



$D \rightarrow$ *the* (0.6) | *those* (0.3)

$N \rightarrow$ *cat(s)* (0.3) | *dog(s)* (0.7)

$V \rightarrow$ *hear(s)* (0.9) | *dog(s)* (0.1)

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow D\ N$

Dynamic programming

$D \rightarrow$ *the* (0.6) | *those* (0.3)

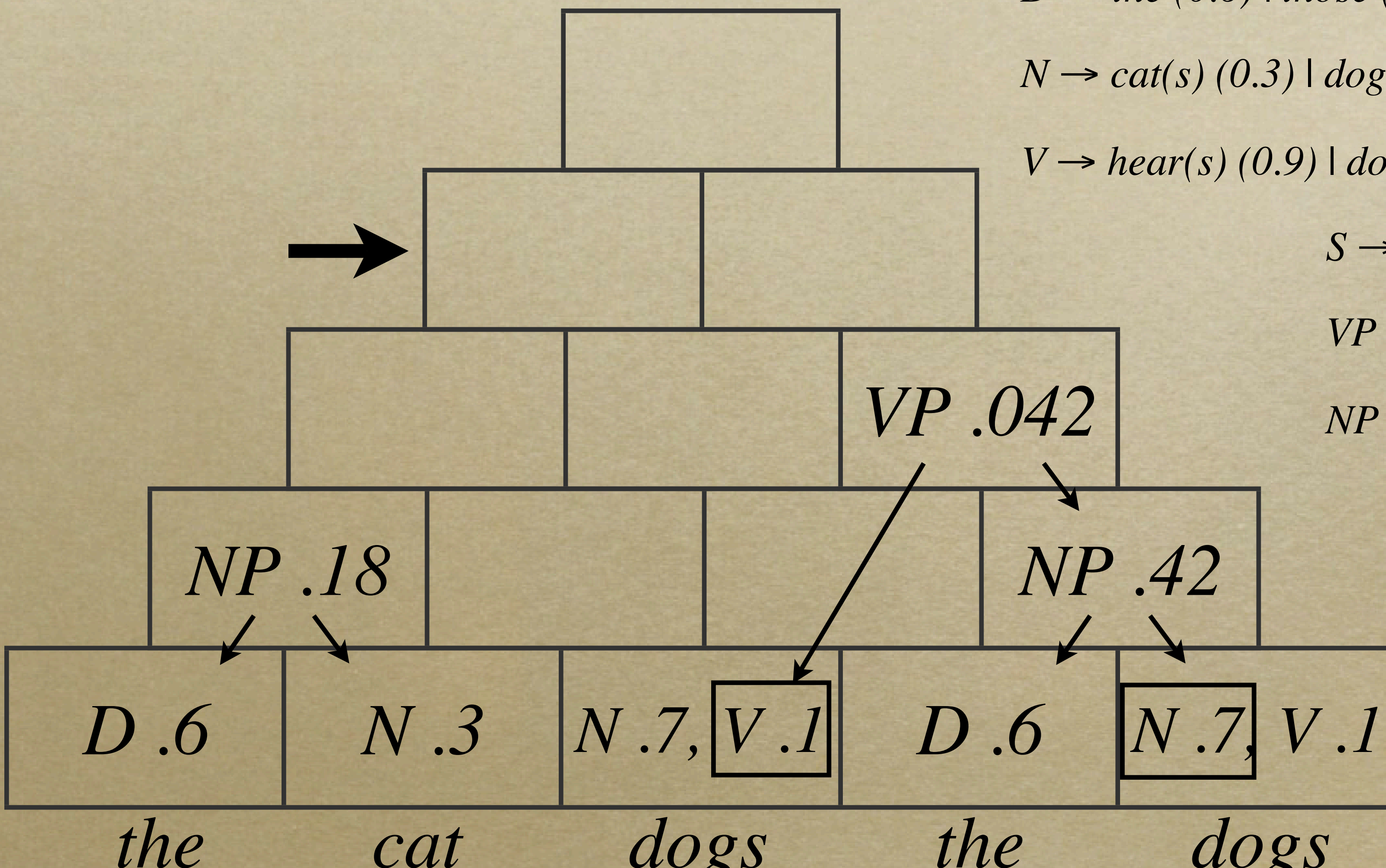
$N \rightarrow$ *cat(s)* (0.3) | *dog(s)* (0.7)

$V \rightarrow$ *hear(s)* (0.9) | *dog(s)* (0.1)

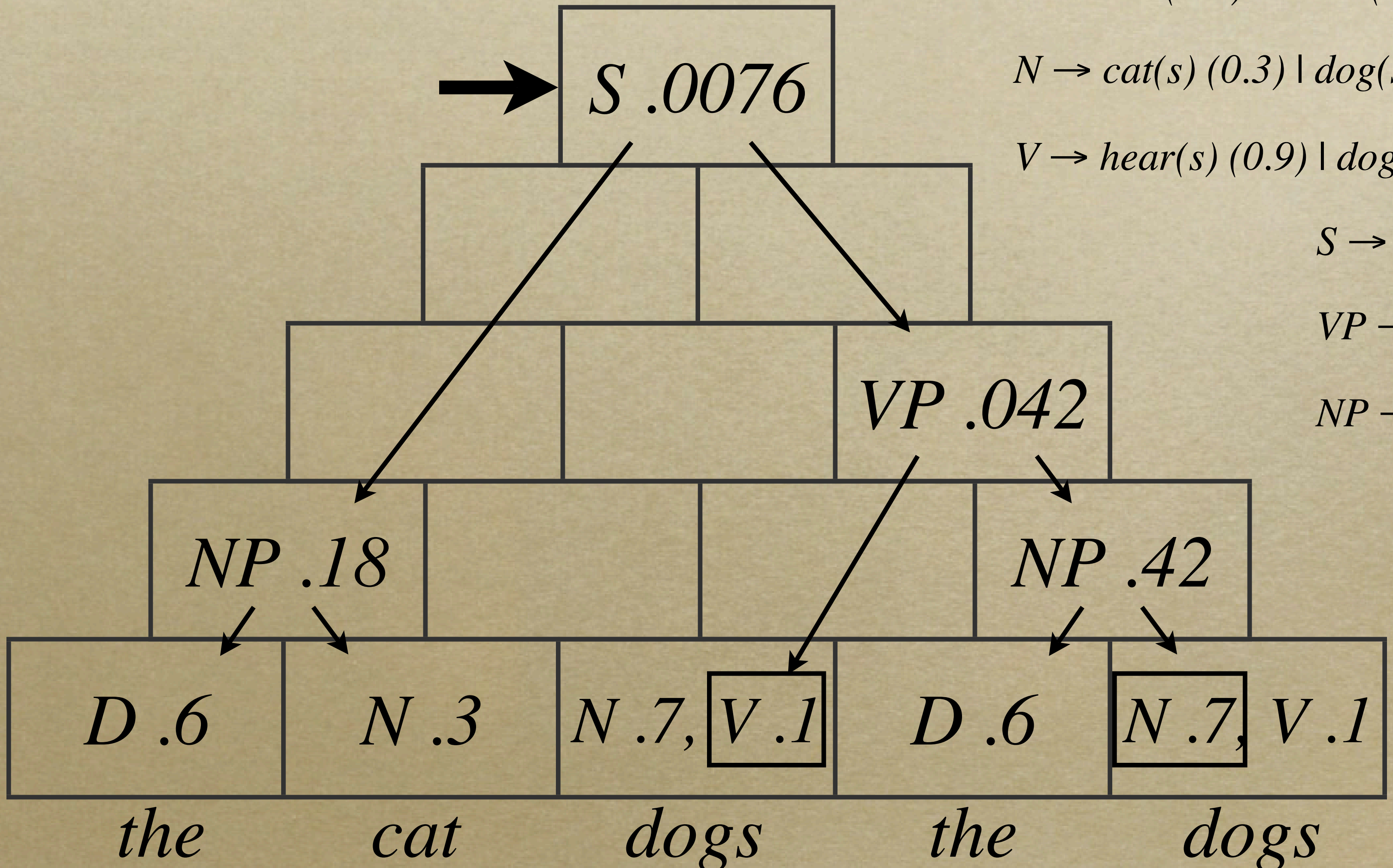
$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$

$NP \rightarrow D\ N$



Dynamic programming



D → the (0.6) | those (0.3)

N → cat(s) (0.3) | dog(s) (0.7)

V → hear(s) (0.9) | dog(s) (0.1)

S → NP VP

VP → V NP

NP → D N

More Powerful Models

- *Link words to their arguments:* **lexicalization**
- *Smooth model for better generalization*
- *Train models using more sophisticated machine learning*

Applications

- *Machine translation*
- *Speech recognition, synthesis, dialog*
- *Information Retrieval*
- *Question Answering*
- *Sentiment Analysis*
- *Spelling/Grammar Checking*
- *Digitization (Optical Character Recognition)*
- *Natural Language Interfaces*
- *Language Education*

Example: Statistical Translation

- *Input: Chinese sentence*
- *Output: English translation*
- *Evaluation: how close is output to a reference translation? (controversial how to measure this!)*
- *Before 1990: write rules by hand.*

Example: Statistical Translation

- *Predominant approach now:* **learn** *to translate from a* **parallel corpus** *of examples.*
  - *Parliamentary proceedings from bilingual countries (Canada, Hong Kong) or the EU or UN; also laws*
  - *News from agencies that publish in multiple languages*
  - *Nowadays: tens-to-hundreds of millions of words each side*

Translation by Modeling

- *Warren Weaver (1948):*

This Russian document is actually an **encoded** *English document! That is, the writer was thinking in English, and somehow the message was garbled into this strange "Russian" stuff. All we have to do is* **decode**!

- *Modern MT: model the source (English sentences) and the channel (translation):*

$$\hat{\mathbf{e}}(\mathbf{c}) \leftarrow \underset{\mathbf{e}}{\arg\max}\, p(\mathbf{e} \mid \mathbf{c}) = \underset{\mathbf{e}}{\arg\max} \frac{p(\mathbf{c} \mid \mathbf{e}) \cdot p(\mathbf{e})}{p(\mathbf{c})} = \underset{\mathbf{e}}{\arg\max}\, p(\mathbf{c} \mid \mathbf{e}) \cdot p(\mathbf{e})$$

Three Statistical MT Problems

- *Build a* **language** *model over English sentences*
  - *Learn from English data!*
- *Build a* **translation** *model that turns English into Chinese*
  - *Learn from parallel data!*
- *Build a* **decoder** *that finds the best English sentence, given Chinese input.*
  - *NP hard for many models!*
  - *Difficult* **search** *problem.*

Translational Structure

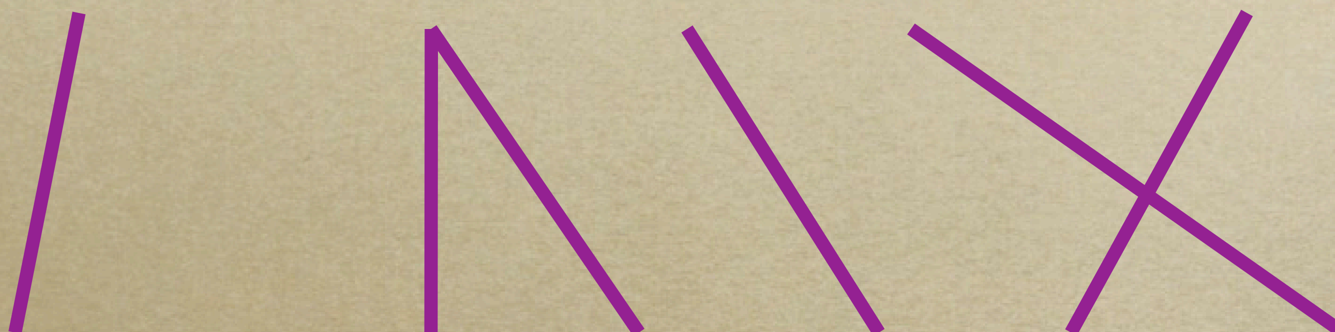
Klimatizovaná jídelna, světlá místnost pro snídaně.

?

Air-conditioned dining room, well-lit breakfast room.

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.

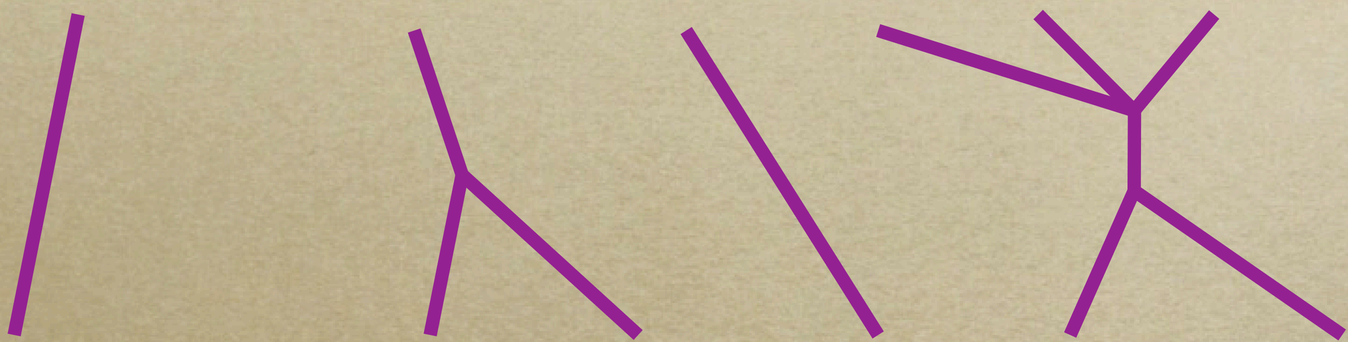


Air-conditioned dining room, well-lit breakfast room.

Word-to-word correspondences?

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.



Air-conditioned dining room, well-lit breakfast room.

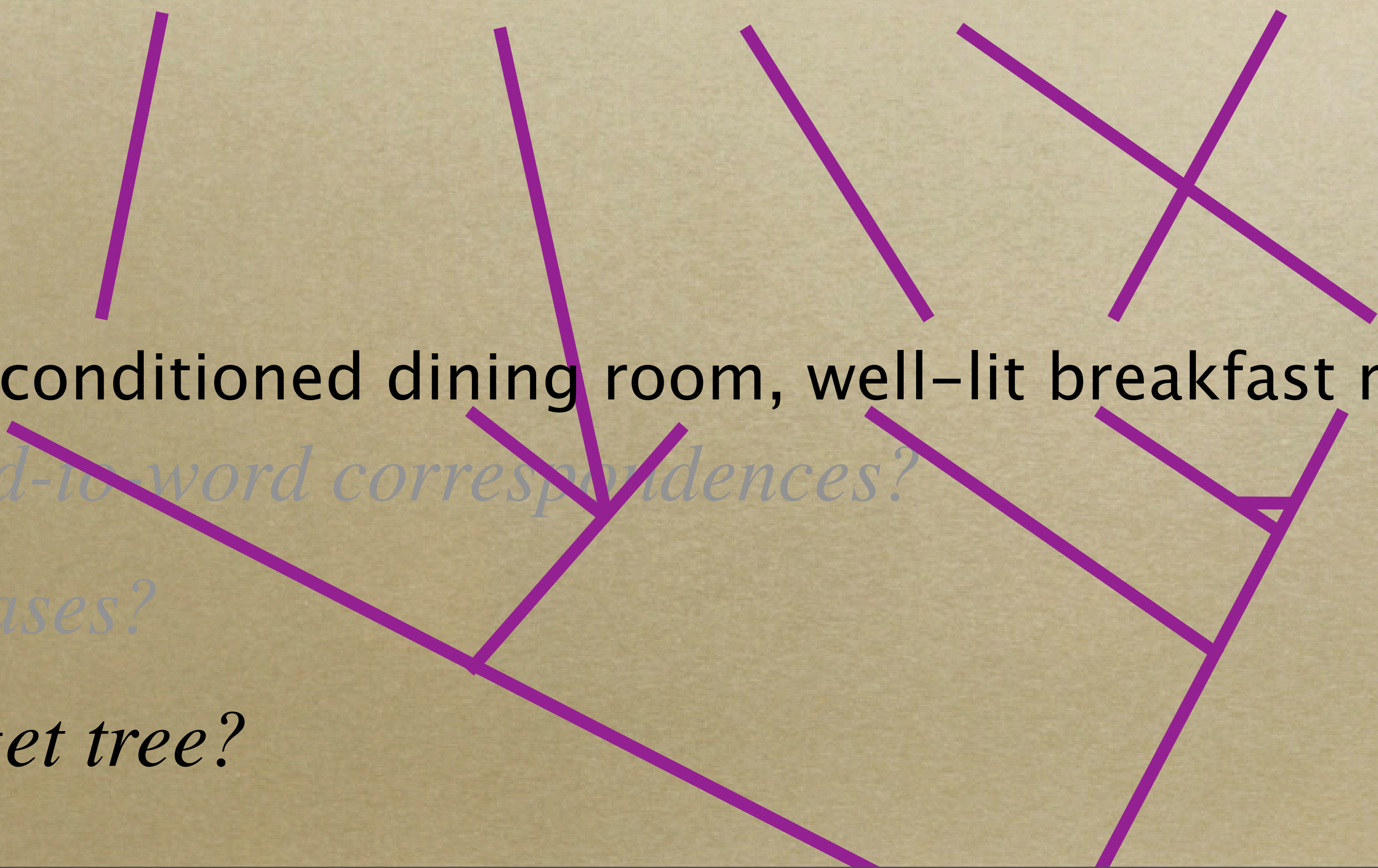
Word-to-word correspondences?

Phrases?

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.

Air-conditioned dining room, well-lit breakfast room.



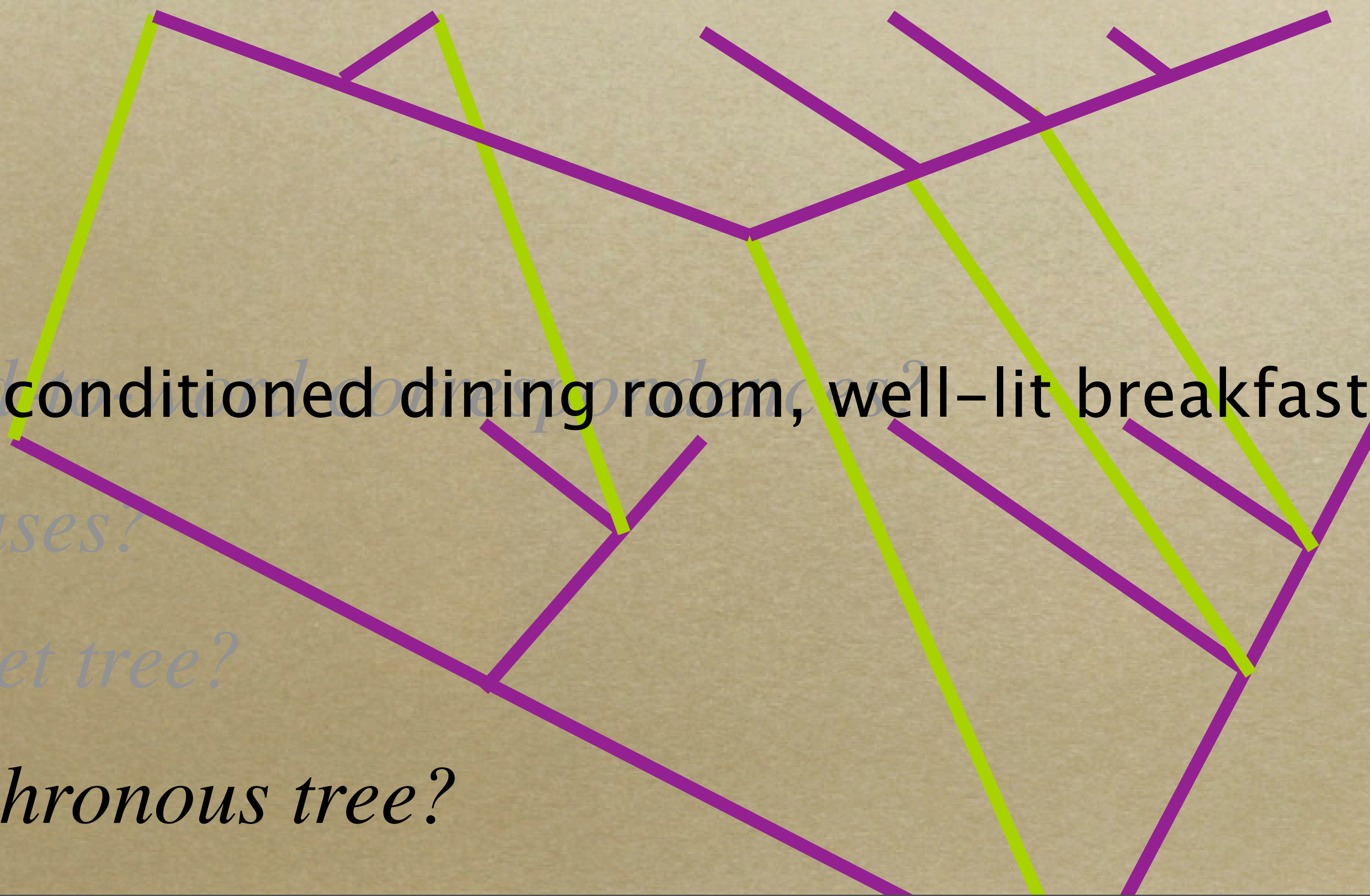
Word-to-word correspondences?

Phrases?

Target tree?

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.



Air-conditioned dining room, well-lit breakfast room.

Word-to-word correspondence?

Phrases?

Target tree?

Synchronous tree?

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.

Air-conditioned dining room, well-lit breakfast room.

Word-to-word correspondences?

Phrases?

Source tree?

Synchronous tree?

Synchronous dependency tree?

Translational Structure

Klimatizovaná jídelna, světlá místnost pro snídaně.

Klimatizovan- jídelna, světl- snídan- místnost

Air-conditioned dining room, well-lit breakfast room.

Word-to-word correspondences?

Phrases?

Source tree?

Synchronous tree?

Synchronous dependency tree?

Czech-prime?

Statistical MT

- *As before:*
  - *Ambiguity at all levels*
  - *Tradeoffs*
- *Will not discuss the models or how they are learned, but lots of interesting stuff here…*

Current Hot Areas

- *Domain: most text ain't newstext!*
  - *Biomedical text*
  - *Blogs*
  - *Conversation*
- *Multilingual NLP: most languages are not like English!*
- *Models and representations (e.g., features) for deeper understanding (e.g., sentiment) or simply better accuracy*
- *Learning from* **unannotated** *data (or less-annotated data, or less annotated data)*
- *How should we evaluate NLP systems?*

Summary

- *Language is hard because it is* **productive** *and* **ambiguous** *on all levels.*
- *NLP is about trading off between*
  - *Accuracy and coverage*
  - *Speed and expressive power*
  - *Human and computational effort*
  - *General mathematical formalisms and specific applications*
- **Optimization**, **search**, *and* **probabilistic inference** *methods underlie much of modern NLP (e.g., dynamic programming, training and applying models).*

Courses of Interest (11nnn)

- *Language and Statistics (I and II!)*
- *Algorithms in NLP*
- *Grammars and Lexicons*
- *Information Extraction*
- *Information Retrieval*
- *Machine Translation*
- *Speech Recognition and Understanding*

11-762 (Noah Smith)

Also 11-411 (u-grad / masters)