

# Language Technology

CUNY Graduate Center, Fall 2014

Professor Liang Huang

[huang@cs.qc.cuny.edu](mailto:huang@cs.qc.cuny.edu)

<http://acl.cs.qc.edu/~lhuang/teaching/nlp>

# Logistics

- Lectures T 11:45-1:45 pm Room 6496
- Personnel
  - Instructor Prof. Liang Huang [huang@cs.qc.cuny.edu](mailto:huang@cs.qc.cuny.edu)
  - TA James Cross [jcross@gc](mailto:jcross@gc)
- Office Hours
  - LH right after class (and occasionally on Fridays)
  - additional office hours available before quizzes/exams
- Homepage: <http://acl.cs.qc.edu/~lhuang/teaching/nlp>



# Doesn't Google know everything?

What animal does a cat eat?

What animal does a cat eat

About 10,700,000 results (0.24 seconds)

[Advanced search](#)

## [What wild animal would eat a domestic cat or dog? - Yahoo!7 Answers](#)

Aug 21, 2008 ... I went outside to get my kitty in a few minutes ago and couldn't ... I hope it's not what you think, I really **do**. But there are wild **animals** ...

[au.answers.yahoo.com](#) › [Pets](#) › [Cats](#) - [Cached](#) - [Similar](#)

[What If my cat eats rawhide?](#) - 2 answers - Aug 14, 2010

[Why does my cat eat grass?](#) - 15 answers - Nov 2, 2008

[More results from au.answers.yahoo.com](#) »

## [WikiAnswers - What animal eats cat](#)

Wild **Animals** question: **What animal eats cat?** Mountain lions **do**, as will coyotes. I've heard of bobcats eating cats too, but i'm not sure it's true.

[wiki.answers.com/Q/What\\_animal\\_eats\\_cat](#) - [Cached](#) - [Similar](#)

Retrieved August 2010

# Even Key Word Queries

- Paris Hilton -- not easy to book! (vs. Boston Hilton)



# Ambiguity

Where can I spot a snow leopard?



Where can I spot a snow leopard

About 27,400,000 results (0.19 seconds)

[Apple - Mac OS X Snow Leopard - The world's most advanced OS](#)

To advance Mac OS X **Leopard**, Apple engineers went deep into the code to streamline, secure, and add new core technologies.

[Buy Mac OS X Snow Leopard now. - Compatibility - Desktop](#)

[www.apple.com/macosx/ - Cached - Similar](#)

[Snow Leopard - Wikipedia, the free encyclopedia](#)

The **snow leopard** (*Uncia uncia*) is a moderately large cat native to the ... their body with small **spots** of the same color on their heads and larger **spots** on ...

[Taxonomy - Etymology - Distribution - Ecology and behaviour](#)

[en.wikipedia.org/wiki/Snow\\_Leopard - Cached - Similar](#)

[Mac OS X Snow Leopard - Wikipedia, the free encyclopedia](#)

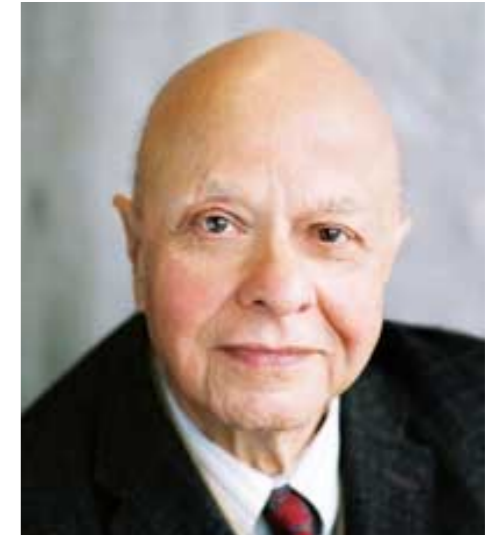
Mac OS X **Snow Leopard** (version 10.6) is the seventh and current major release of Mac OS X, Apple's desktop and server certified Unix operating system. ...

[en.wikipedia.org/wiki/Mac\\_OS\\_X\\_Snow\\_Leopard - Cached - Similar](#)

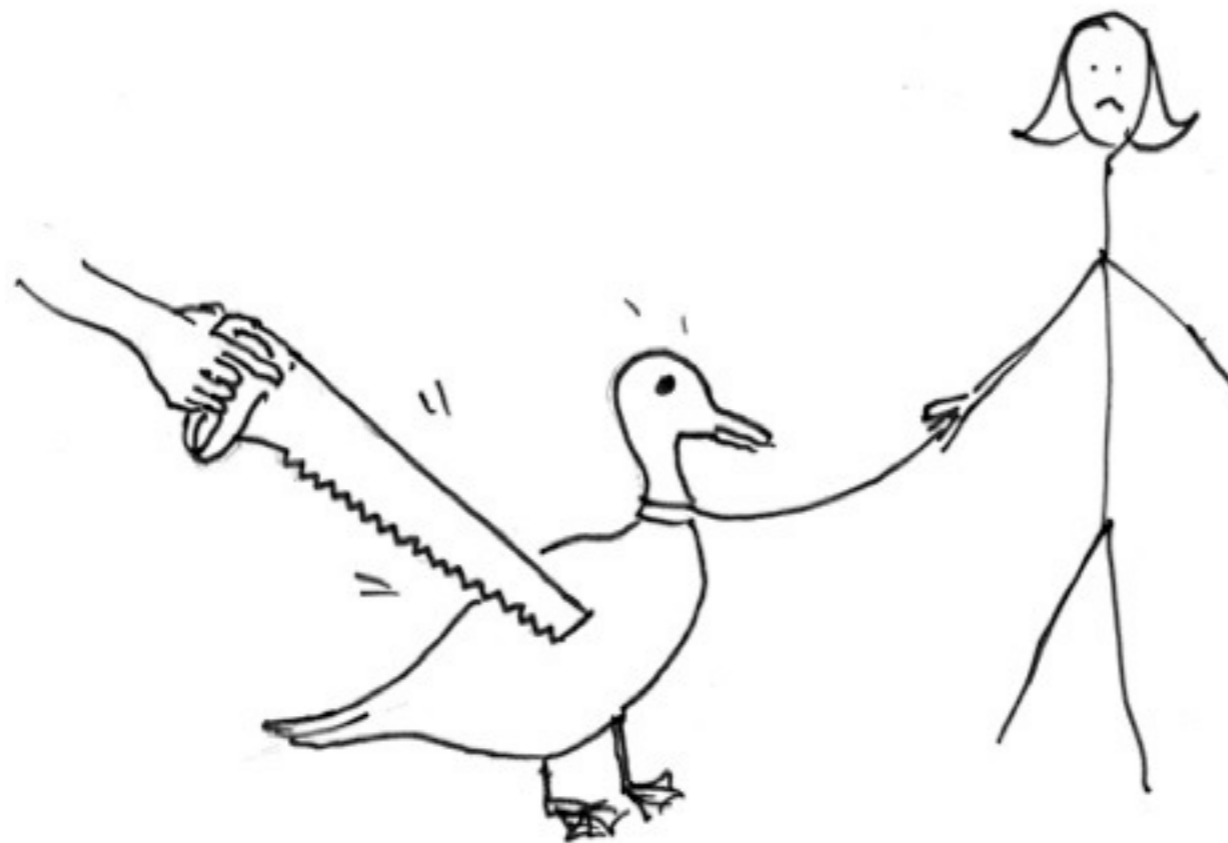
# More about Ambiguities

- to middle school kids: what does this sentence mean?

I saw her duck.



Aravind Joshi

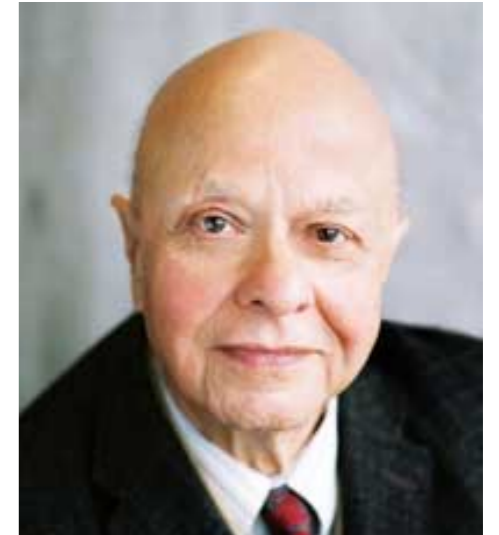


lexical ambiguity  
(word-sense)

# More about Ambiguities

- to middle school kids: what does this sentence mean?

I eat sushi with tuna.



Aravind Joshi

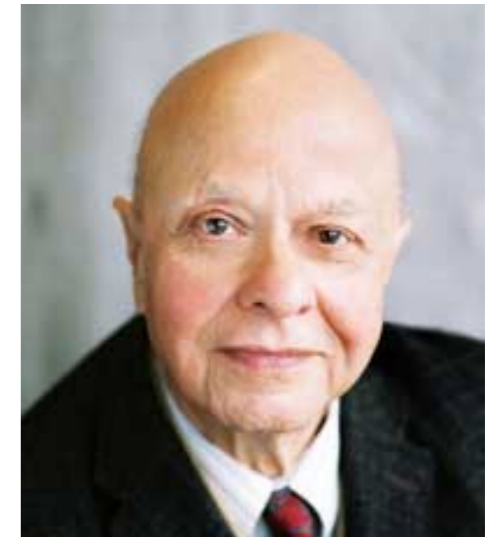


structural ambiguity  
(PP-attachment)

# More about Ambiguities

- to middle school kids: what does this sentence mean?

I eat sushi with tuna.



Aravind Joshi



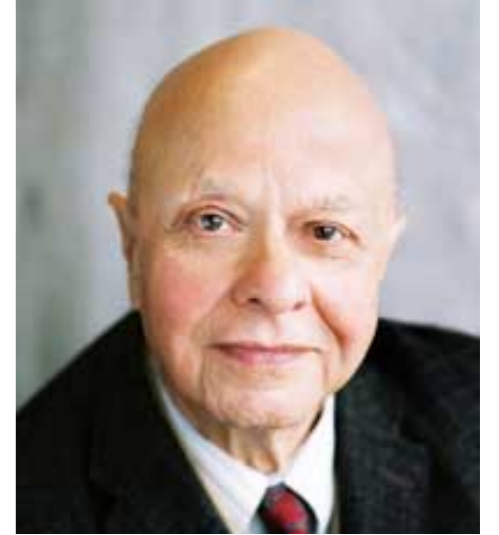
lexical ambiguity  
(word-sense)



# More about Ambiguities

- to middle school kids: what does this sentence mean?

Everybody loves somebody.



Aravind Joshi



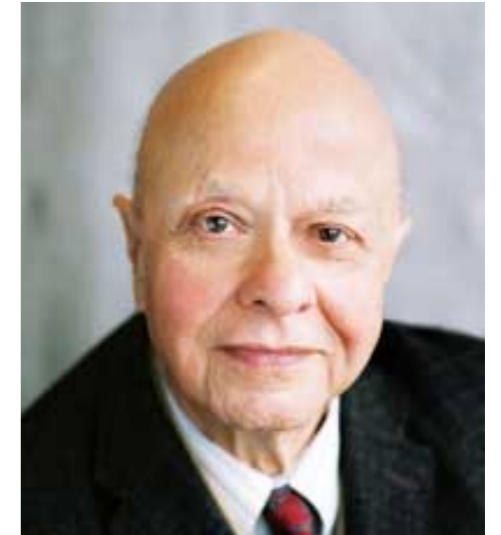
???

structural ambiguity  
(quantifier scope)

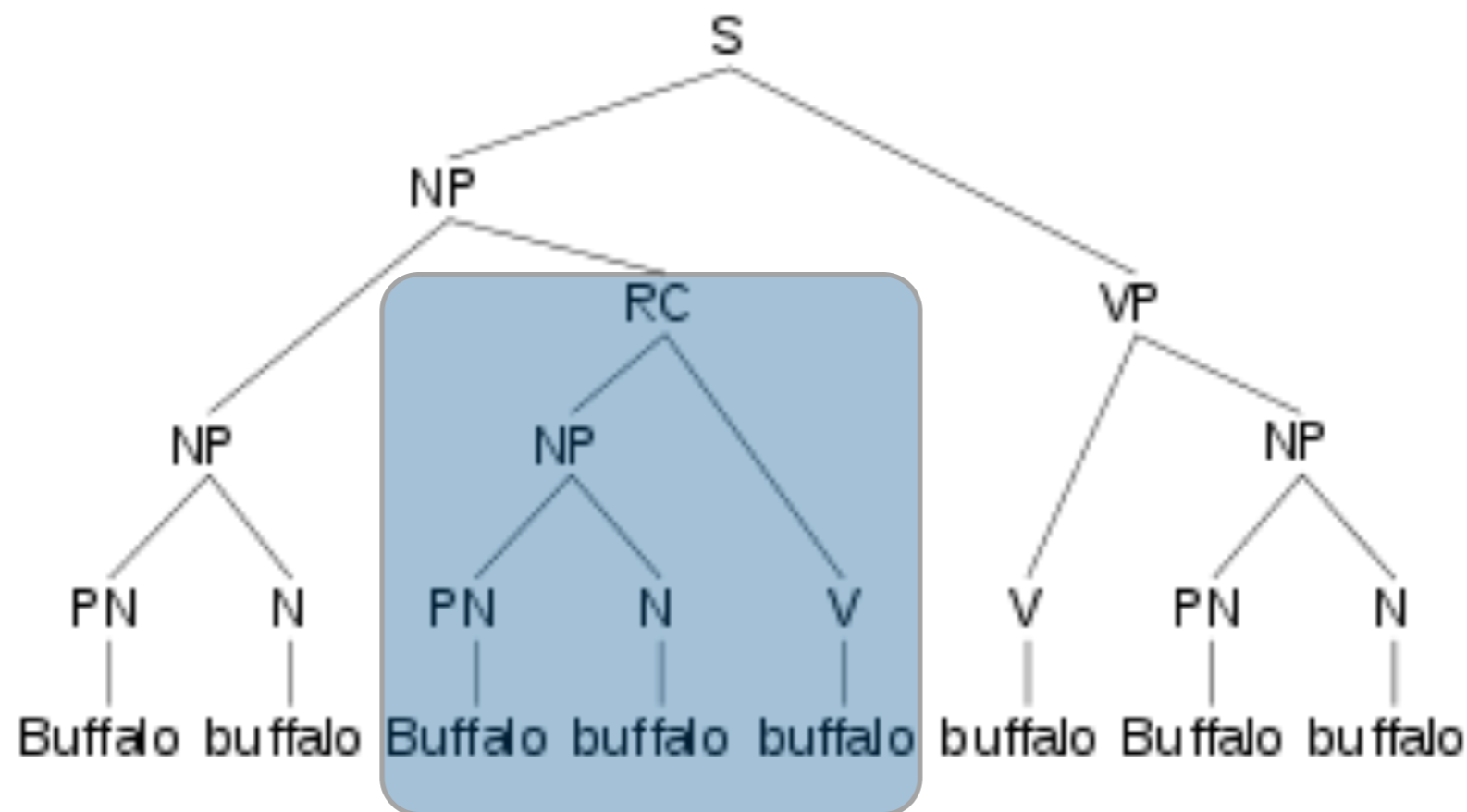
# More about Ambiguities

- to middle school kids: what does this sentence mean?

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo



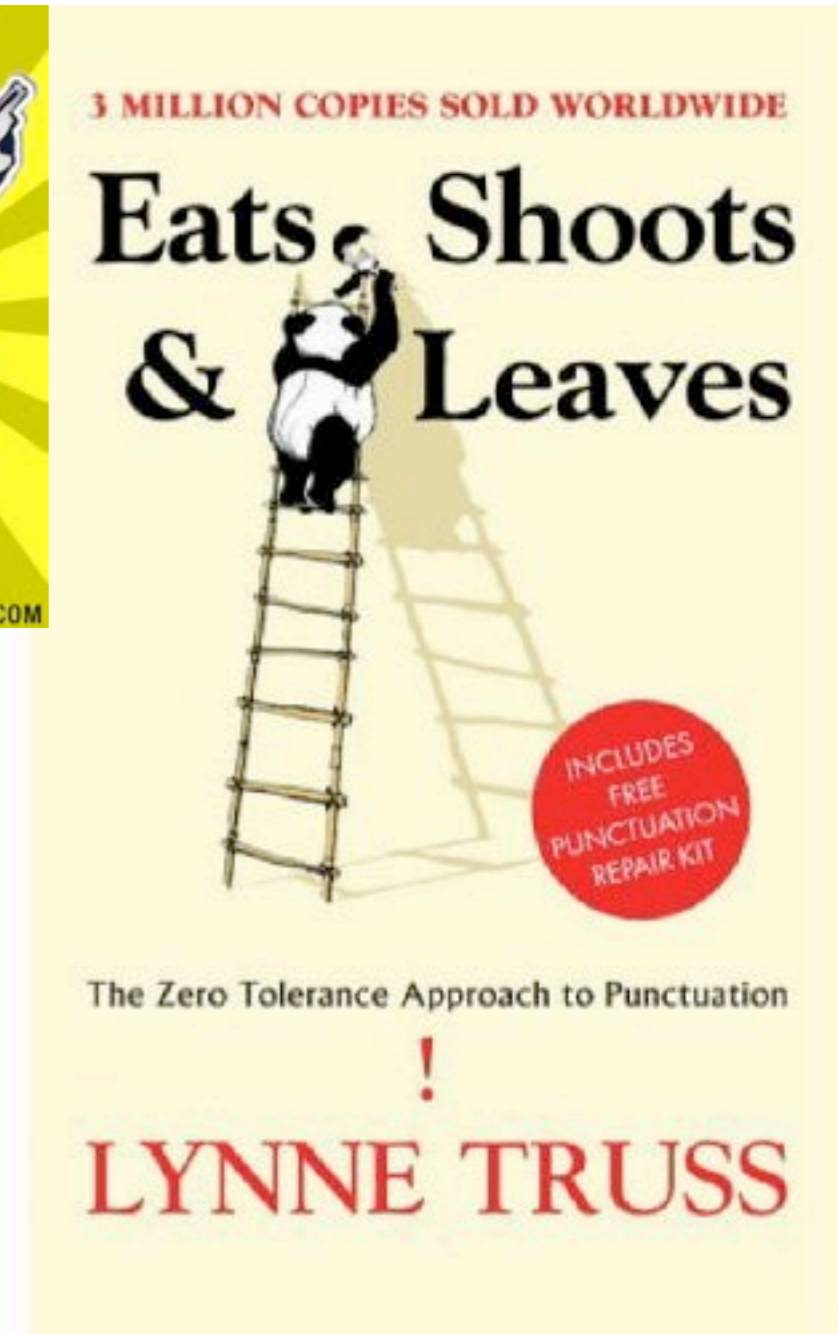
Aravind Joshi



Dogs dogs dog dog dogs.  
Police police police police police

# Prosody and Ambiguity

- a panda
  - eats [shoots]<sub>N</sub> and [leaves]<sub>N</sub>
  - eats [shoots]<sub>V</sub> and [leaves]<sub>V</sub>
- prosody marks this ambiguity by
  - prominence on *eats*
  - break between *eats* and *shoots*



# Ambiguities in Translation



zi zhu zhong duan  
自 助 终 端

self help terminal device

# Ambiguities in Translation



Google translate: carefully slide



# If you are stolen...

一旦失窃要报警，切莫姑息又养奸

**If you are stolen, call the police at once.**

Google translate: Once the theft to the police



上海市公安局城市轨道交通分局

Urban Mass Transportation Branch Shanghai Public Security Bureau



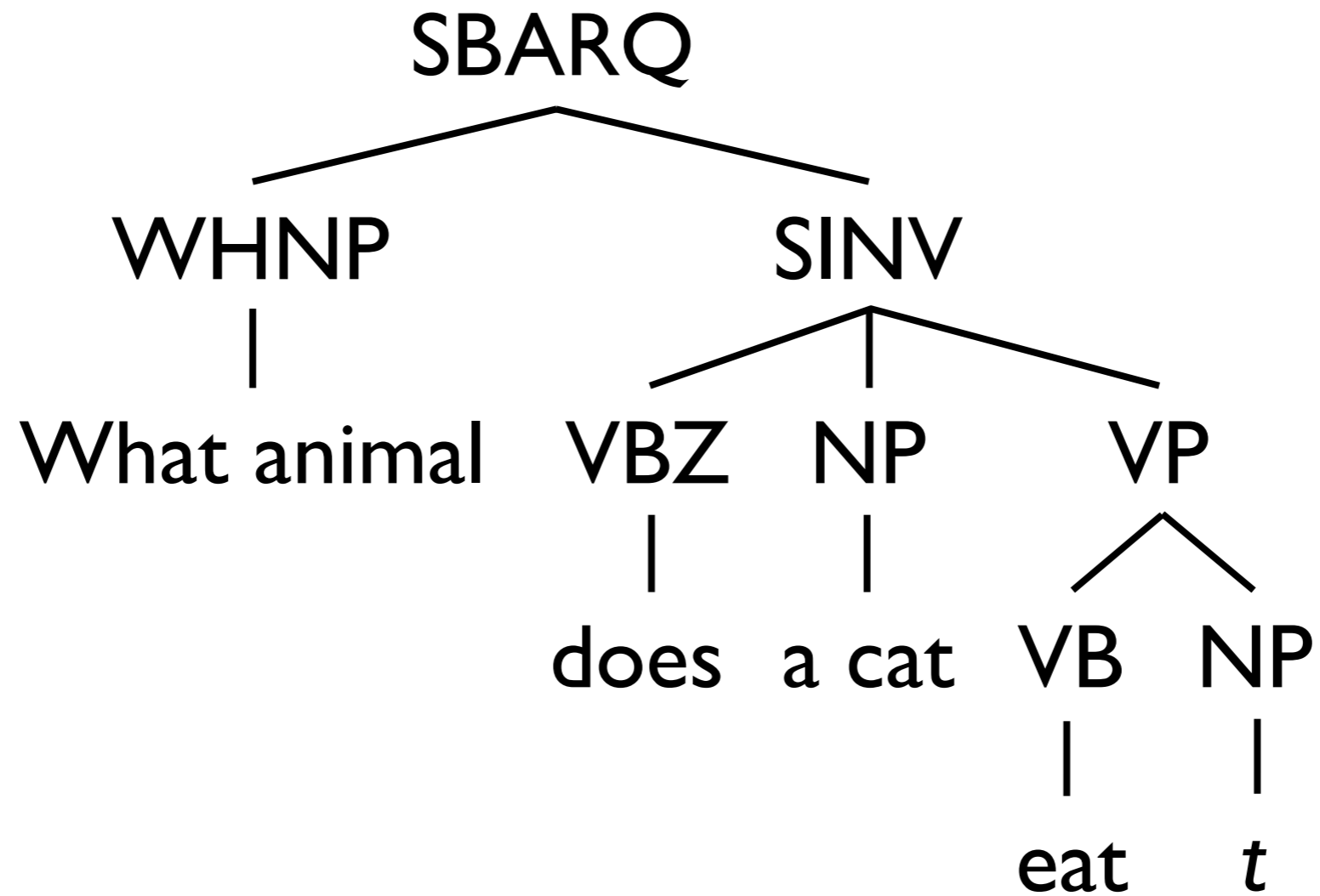
ENGLISH FUNNY.com

or even...



clear evidence that NLP is used in real life!

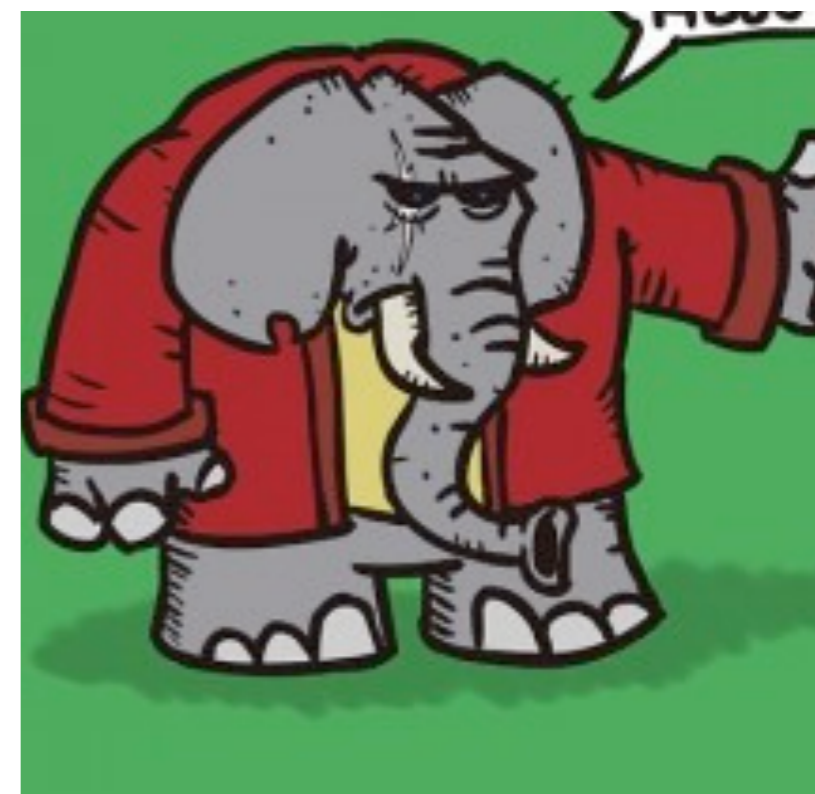
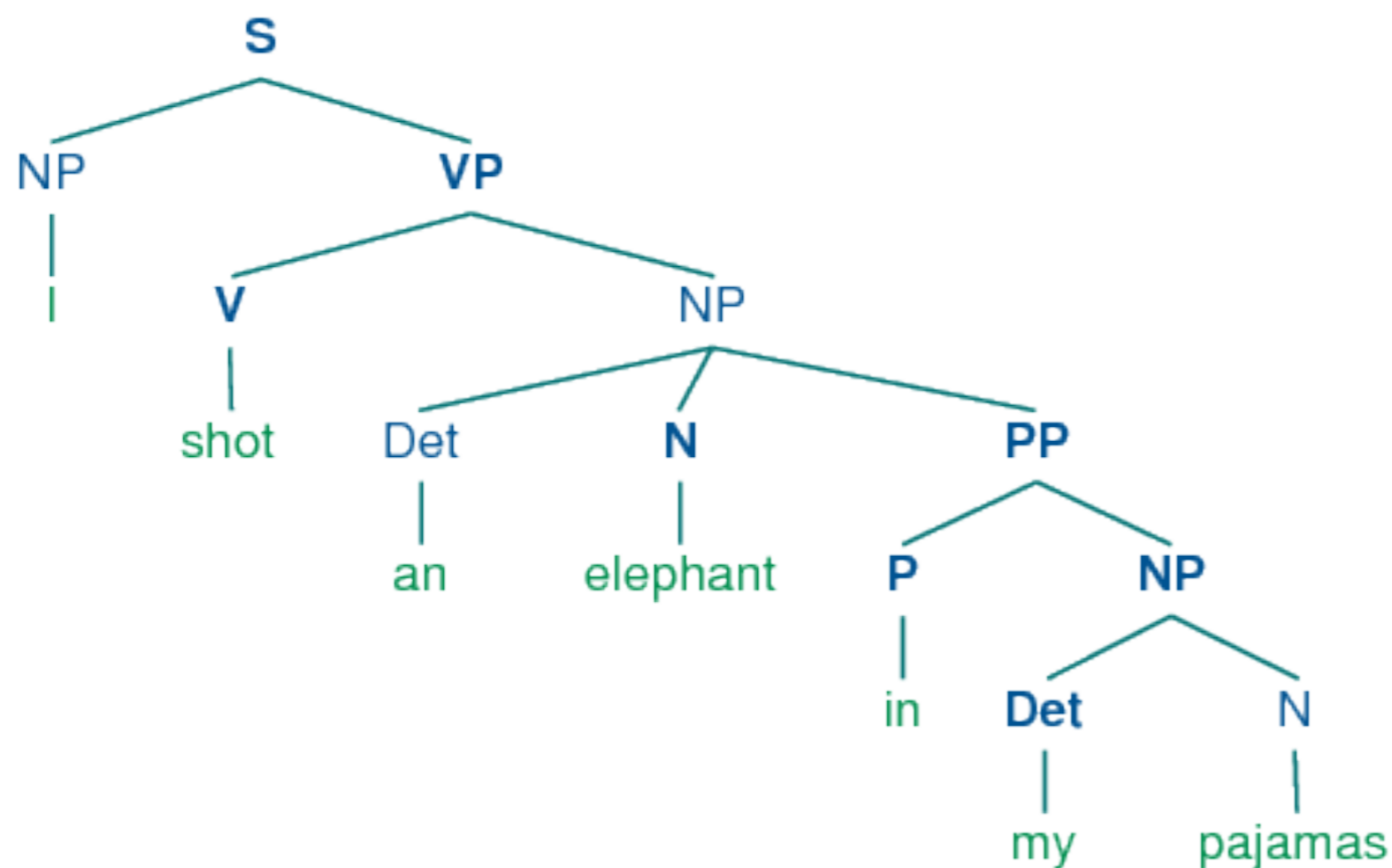
# Grammar





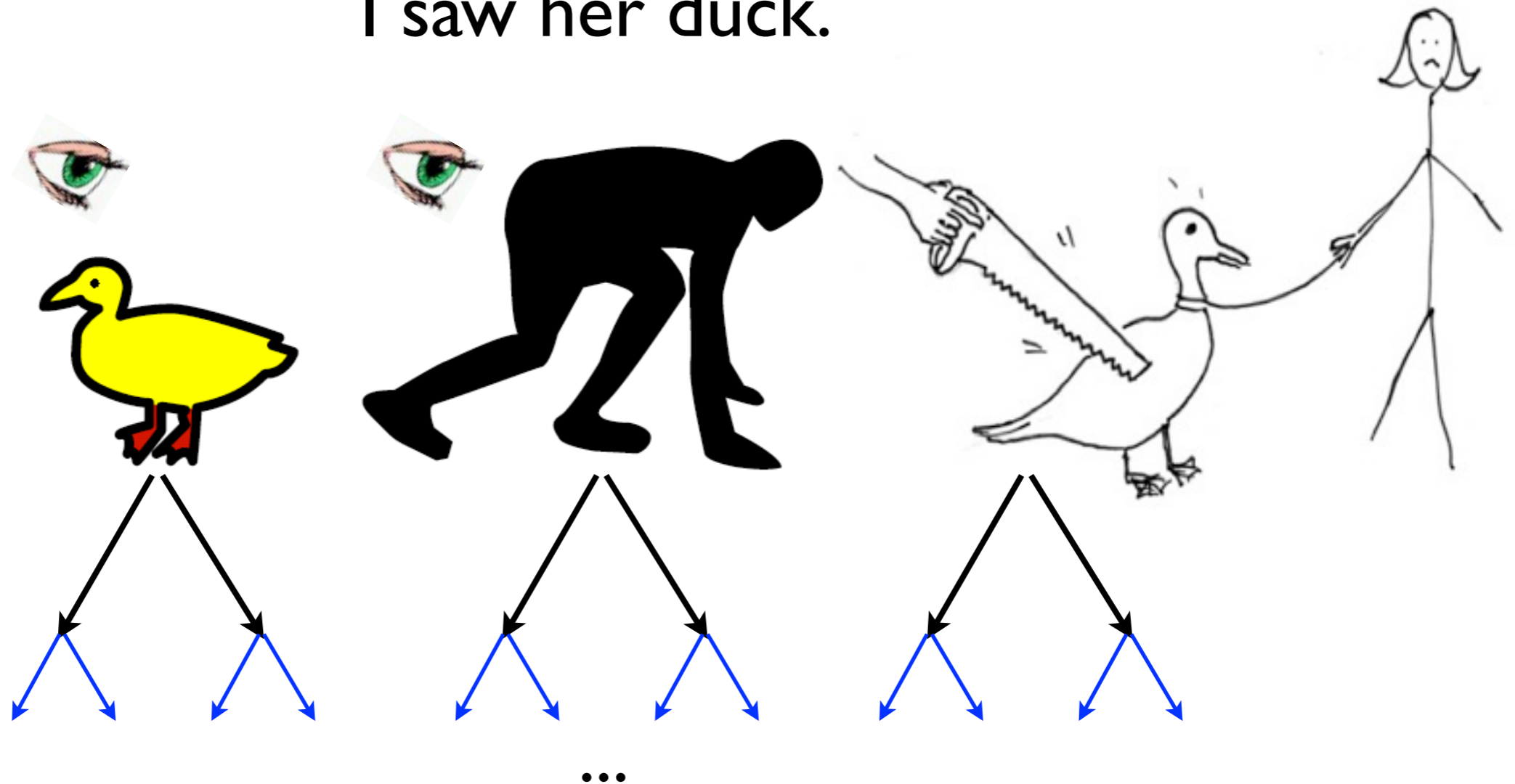
# PP Attachment Ambiguity

One morning in Africa,  
I shot an elephant in my pajamas;  
how he got into my pajamas I'll never know.



# Ambiguity Explosion

I saw her duck.

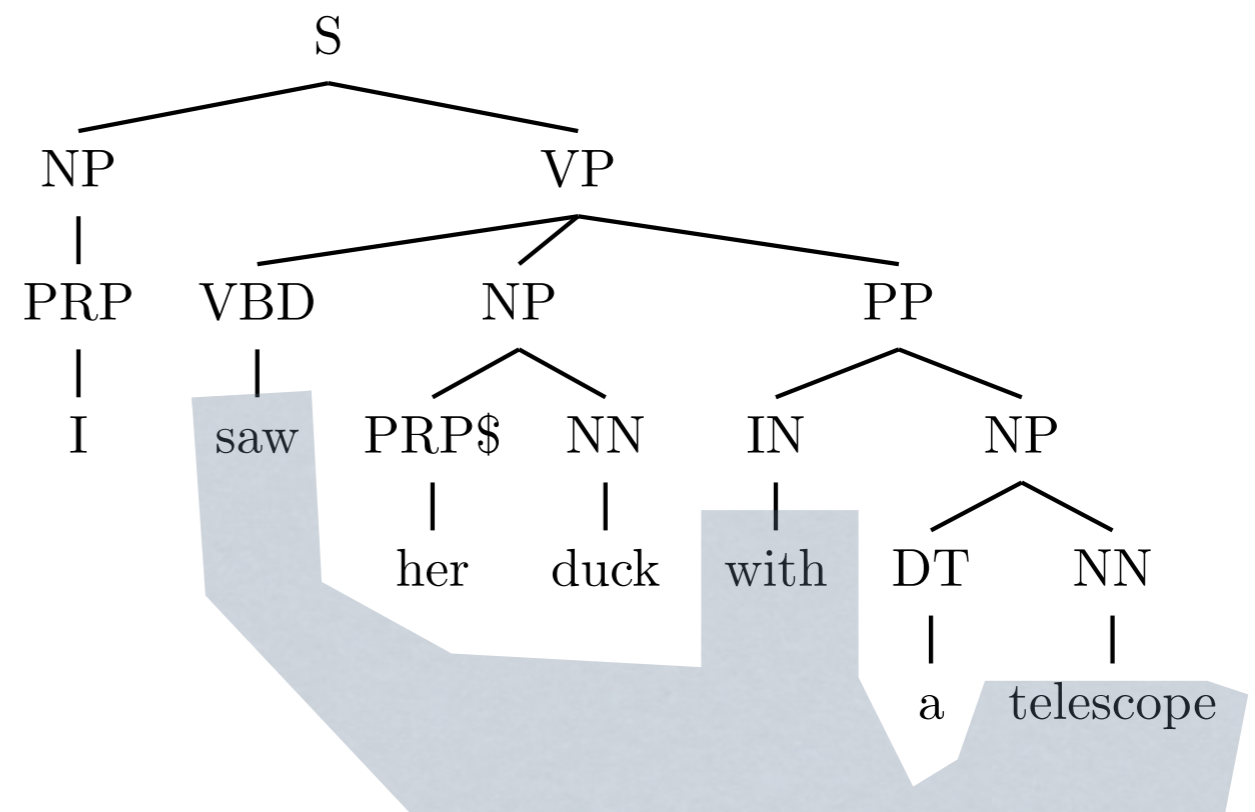
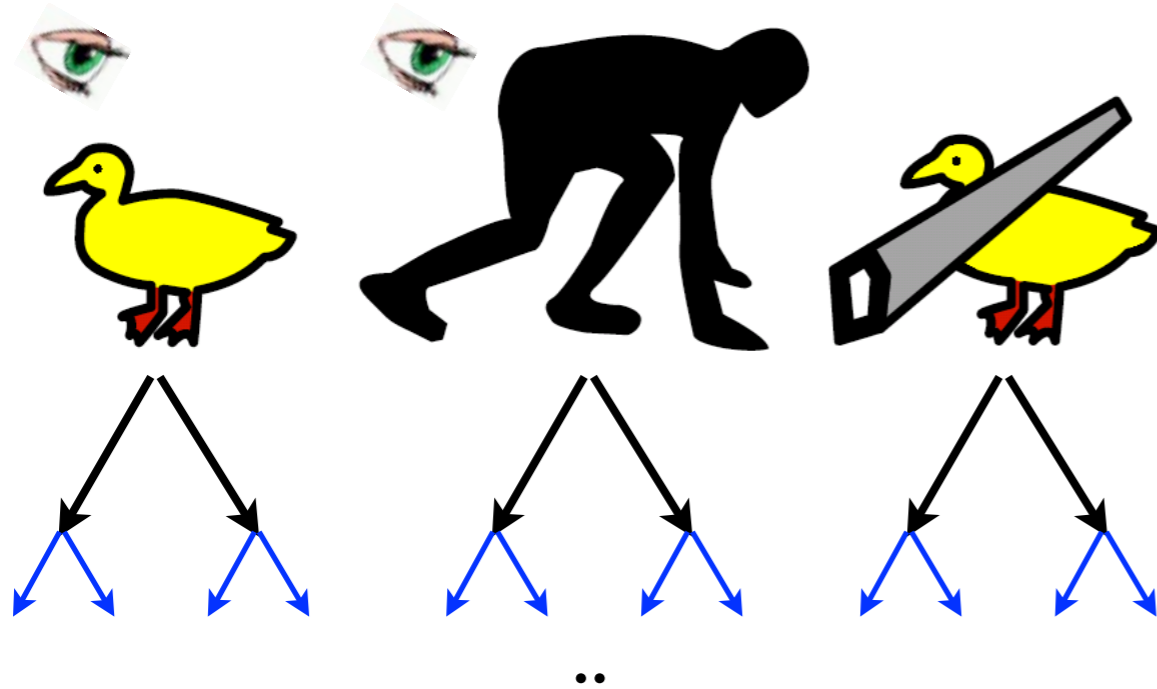


- how about...

- I saw her duck with a telescope.
- I saw her duck with a telescope in the garden...

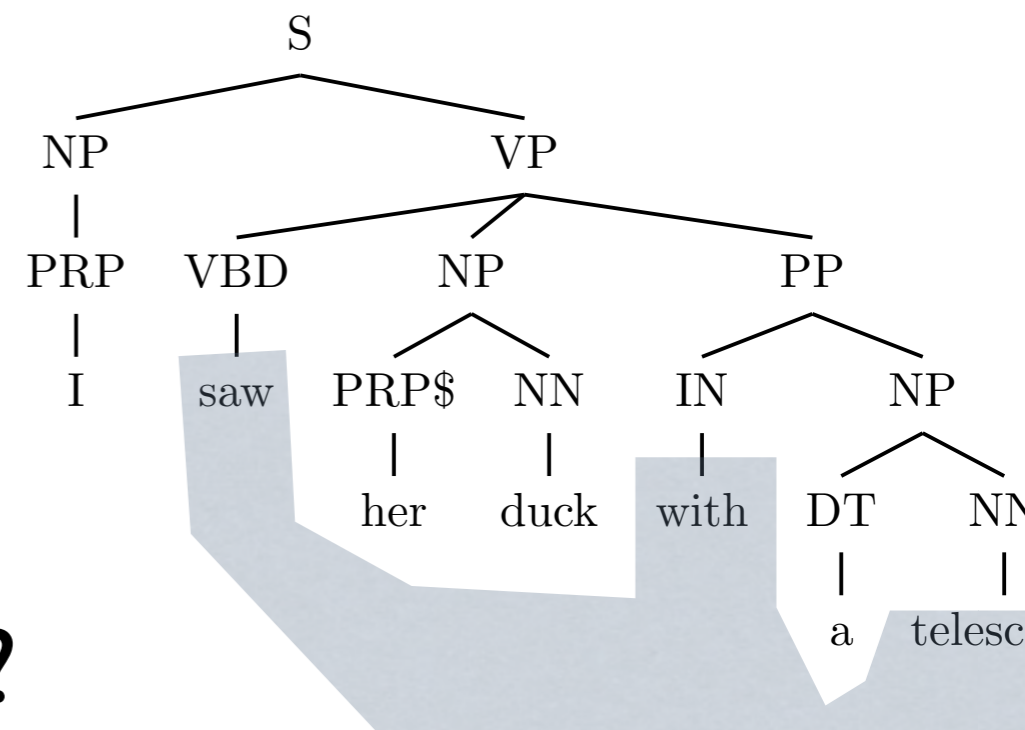
# Ambiguity Explosion

- exponential explosion of the search space
  - Q1: how to *represent* ambiguities (compactly)?
  - Q2: how to *search* over this space (efficiently)?
  - Q3: how to *rank* different hypotheses?



# Answers...

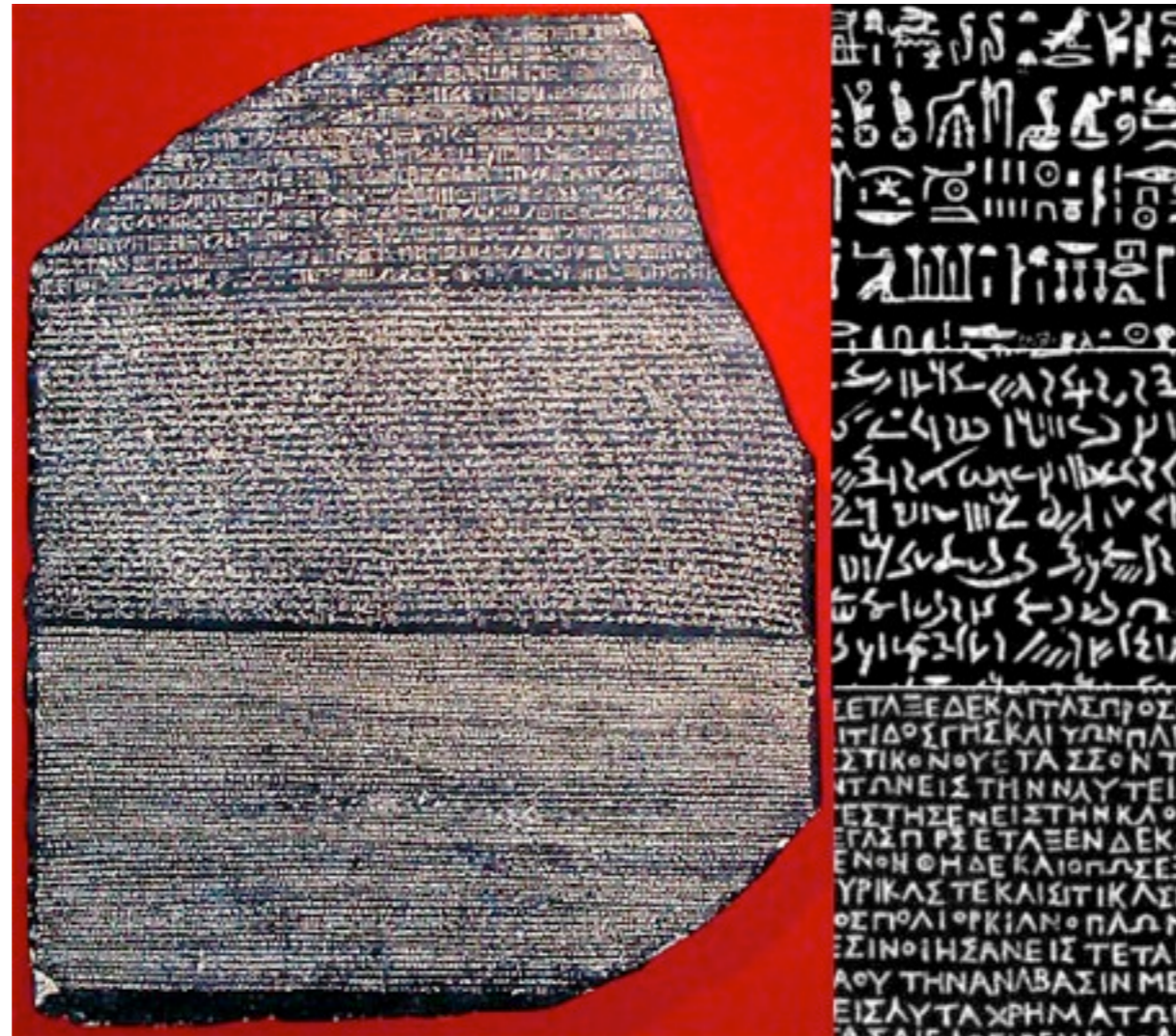
- Q1: how to *represent* ambiguities?
  - context-free grammar (unit 2)
  - finite-state automata (unit 1)
- Q2: how to *search* in this space?
  - dynamic programming (units 1&2)
- Q3: how to *rank* these hypotheses?
  - weighted grammar (units 1-3)
  - weights *learned* from data
    - (saw, with, telescope) seen more often in texts



# Why Learning?

- learning is better than hand-written rules, because:
  - less work; easily adapts to new languages/domains
    - Powerset (now bing.com): 15 years for English grammar!
    - now they are writing their Chinese grammar...
    - and languages constantly change!
  - learning *can* work, and often works better!
    - machine translation: used to be dominated by rule-based
      - now statistical methods are better: google vs. systran
      - google learns from the web, and translates 40+ langs

# Example - Rosetta Stone



- the most famous (tri-)parallel text
- machines can do the same job! (if given parallel text)
  - UN/EU/Ca proceedings, News, tech manuals, ...

# A sci-fi example

(Knight, 1997)

Your assignment: translate this Centauri  
sentence into Arcturan

farok crrrok hihok yorok klok kantok ok-yurp

# farok crrrok hihok yorok klok kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok klok .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok crrrok hihok yorok zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>



# farok crrrok hihok yorok klok kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok klok .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok crrrok hihok yorok zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>

# farok **crrrok** hihok yorok klok kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok klok .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok <b>crrrok</b> hihok yorok zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>

# farok crrrok **hihok** yorok klok kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok <b>hihok</b> ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok klok .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok crrrok <b>hihok</b> yorok zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok <b>hihok</b> mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>

# farok crrrok hihok **yorok** klok kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok <b>yorok</b> ghirok klok .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok crrrok hihok <b>yorok</b> zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>

# farok crrrok hihok yorok **clock** kantok ok-yurp

<p>1c. ok-voon ororok sprok .</p> <p>1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok .</p> <p>7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok .</p> <p>2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok .</p> <p>8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok .</p> <p>3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp .</p> <p>9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok .</p> <p>4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok <b>clock</b> .</p> <p>10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok .</p> <p>5a. totat jjat quat cat .</p>	<p>11c. lalok nok crrrok hihok yorok zanzanok .</p> <p>11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok .</p> <p>6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok .</p> <p>12a. wat nnat forat arrat vat gat .</p>

# farok **crrrok** hihok yorok klok kantok ok-yurp

<p>1c. ok-voon ororok sprok . 1a. at-voon bichat dat .</p>	<p>7c. lalok farok ororok lalok sprok izok enemok . 7a. wat jjat bichat wat dat vat eneat .</p>
<p>2c. ok-drubel ok-voon anak plok sprok . 2a. at-drubel at-voon pippat rrat dat .</p>	<p>8c. lalok brok anak plok nok . 8a. iat lat pippat rrat nnat .</p>
<p>3c. erok sprok izok hihok ghirok . 3a. totat dat arrat vat hilat .</p>	<p>9c. wiwok nok izok kantok ok-yurp . 9a. totat nnat quat oloat at-yurp .</p>
<p>4c. ok-voon anak drok brok jok . 4a. at-voon krat pippat sat lat .</p>	<p>10c. lalok mok nok yorok ghirok klok . 10a. wat nnat gat mat bat hilat .</p>
<p>5c. wiwok farok izok stok . 5a. totat jjat quat cat .</p>	<p>11c. lalok nok <b>crrrok</b> hihok yorok zanzanok . 11a. wat nnat arrat mat zanzanat .</p>
<p>6c. lalok sprok izok jok stok . 6a. wat dat krat quat cat .</p>	<p>12c. lalok rarok nok izok hihok mok . 12a. wat nnat forat arrat vat gat .</p>

# A sci-fi example

(Knight, 1997)

Your assignment: translate this Centauri sentence into Arcturan

farok crrrok hihok yorok klok kantok ok-yurp  
jjat arrat mat bat oloat at-yurp

```
graph TD; C1[farok] --> A1[jjat]; C2[crrrok] --> A2[arrat]; C3[hihok] --> A3[mat]; C4[yorok] --> A4[bat]; C5[klok] --> A5[oloat]; C6[kantok] --> A6[at-yurp]; C7[ok-yurp] --> A6;
```

Are these Arcturan words in Arcturan order?

# Clients do not sell pharmaceuticals in Europe .

<p>1e. Garcia and associates . 1s. Garcia y asociados .</p>	<p>7e. the clients and the associates are enemies . 7s. los clients y los asociados son enemigos .</p>
<p>2e. Carlos Garcia has three associates . 2s. Carlos Garcia tiene tres asociados .</p>	<p>8e. the company has three groups . 8s. la empresa tiene tres grupos .</p>
<p>3e. his associates are not strong . 3s. sus asociados no son fuertes .</p>	<p>9e. its groups are in Europe . 9s. sus grupos estan en Europa .</p>
<p>4e. Garcia has a company also . 4s. Garcia tambien tiene una empresa .</p>	<p>10e. the modern groups sell strong pharmaceuticals . 10s. los grupos modernos venden medicinas fuertes .</p>
<p>5e. its clients are angry . 5s. sus clientes estan enfadados .</p>	<p>11e. the groups do not sell zenzanine . 11s. los grupos no venden zanzanina .</p>
<p>6e. the associates are also angry . 6s. los asociados tambien estan enfadados .</p>	<p>12e. the small groups are not modern . 12s. los grupos pequenos no son modernos .</p>



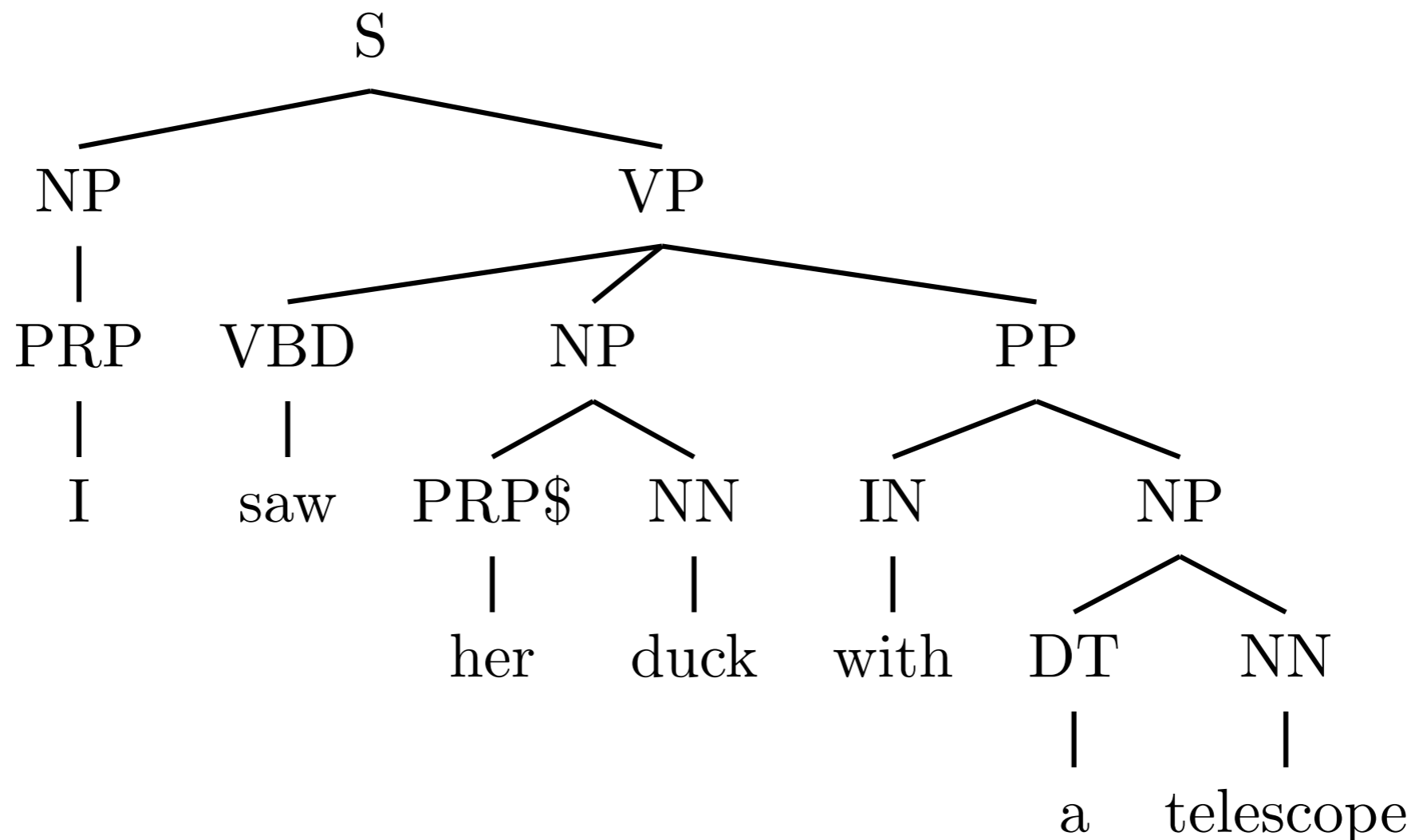
# Take Home Message

- languages are *beyond* just bags of words!
  - ambiguity is everywhere, and NLP is all about that
- we'll teach machines how to read and translate...
  - and how to *learn* to read and translate from data
- have fun in this class! :)



# Basic Linguistic Structures

- parse tree; grammar rules like  $S \rightarrow NP VP$ ;  $NP \rightarrow PRP$
- nonterminals like  $S$ ,  $NP$ ,  $VP$ , ...
- preterminals (part-of-speech tags):  $PRP$ ,  $VBD$ ,  $IN$



# Part-of-Speech Tags

- Penn Treebank Part-of-Speech Tags

CC	the cat <b>and</b> the dog	PRPS	<b>his</b> mat
CD	<b>three</b> cats	RB	The <b>very</b> fat cat
DT	<b>the</b> cat	RBR	The <b>more</b> corpulent cat
EX	<b>there</b> was a cat on the mat	RBS	The <b>most</b> corpulent cat
FW	le <b>chat</b>	RP	The cat sat <b>down</b>
IN	<b>on</b> the mat	SYM	50% of cats sit on mats.
JJ	the <b>fat</b> cat	TO	The cat walked <b>to</b> the mat
JJR	the <b>fatter</b> cat	UH	<b>Hey</b> , the cat sat on the mat
JJS	the <b>fattest</b> cat	VB	The cat wants to <b>sit</b> on the mat
LS	<b>a.</b> the cat	VBD	The cat <b>sat</b> on the mat
MD	the cat <b>might</b> sit on the mat	VBG	The cat is <b>sitting</b> on the mat
NN	the <b>cat</b>	VBN	The cat has <b>sat</b> on the mat
NNS	the <b>cats</b>	VBP	I <b>sit</b> on the mat sometimes
NNP	<b>Felix</b> sat on the mat.	VBZ	The cat <b>sits</b> on the mat
NNPS	Were there any mats in <i>Cats</i> ?	WDT	<b>Which</b> mat does the cat sit on?
PDT	<b>all</b> the cats	WP	<b>Who</b> sits on the mat?
POS	the cat's mat	WPS	<b>Whose</b> mat does he sit on?
PRP	<b>he</b> sat on the mat	WRB	<b>Where</b> does the cat sit?

# Nonterminal Labels

ADJP	the <b>very fat</b> cat
ADVP	The cat sat on the mat <b>very happily</b> .
CONJP	The cat <b>as well as</b> the dog sat on the mat.
FRAG	Who sat on the mat? <b>The cat</b> .
INTJ	<b>Oh no</b> , the cat sat on the mat!
LST	The following sat on the mat: <b>a.</b> the cat <b>b.</b> the dog
NAC/NML	<b>Secretary of Defense Dick Cheney</b> sat on the mat.
NP	<b>the fat cat</b>
NX	the <b>fat cat</b> and <b>skinny dog</b>
PP	on the mat
PRN	The cat ( <b>not the dog</b> ) sat on the mat.
PRT	The cat sat <b>down</b> .
QP	<b>one hundred and one</b> cats sat on the mat
RRC	the mat <b>the cat sat on</b>
S	<b>The cat sat on the mat.</b>
SBAR	The dog wondered <b>why the cat sat on the mat</b> .
SBARQ	<b>Did the cat sit on the mat?</b>
SINV	<b>There sat the cat on the mat.</b>
SQ	Did <b>the cat sit on the mat?</b>
UCP	the <b>fat and growing</b> cat
VP	The cat <b>sat on the mat</b> .
WHADJP	<b>How fat</b> was the cat?
WHADVP	<b>How long</b> did the cat sit?
WHNP	<b>Which mat</b> did the cat sit on?
WHPP	<b>On which mat</b> did the cat sit?
X	<b>The fatter</b> the cat, <b>the bigger</b> the mat.