Natural Language Processing

Regina Barzilay

What is NLP?

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah...

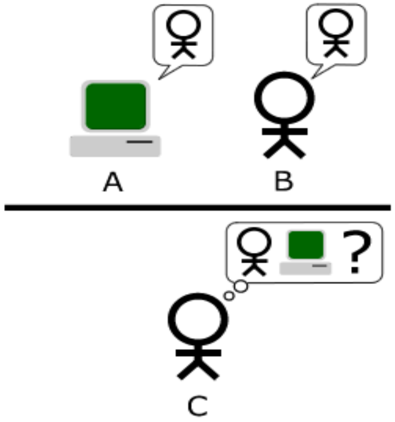
- Goal: intelligent processing of human language
  - Not just effective string matching
- Applications of NLP technology:
  - Less ambitious (but practical goals): spelling corrections, name entity extraction
  - Ambitious goals: machine translations, language-based UI, summarization, question-answering

NLP is AI-complete

All the difficult problems in artificial intelligence manifest themselves in NLP problems

Turing Test: links machine intelligence with the ability to process language

The interrogator C needs to determine which player - A or B - is a computer and which is a human.



Passing Turing Test

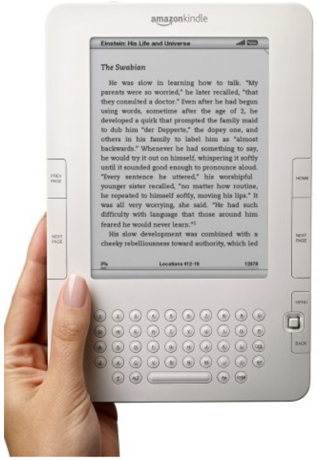
Turing (1950): "I believe that in about fifty years' it will be possible, to programme computers, with a storage capacity of about 109, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning."

ELIZA (Weizenbaum, 1966): first computer dialogue system
based on keyword matching

young woman: Men are all alike.
eliza: In what way?
young woman: They're always bugging us about something specific or other.
eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?

Speech Processing

- Automatic Speech Recognition (ASR):
  - Performance: 0.3% for digital string, 5% for dictation, 50%+TV
- Text to Speech (TTS):
  - Performance: totally intelligible (if sometimes unnatural)



Einstein: His Life and Universe

The Swabian

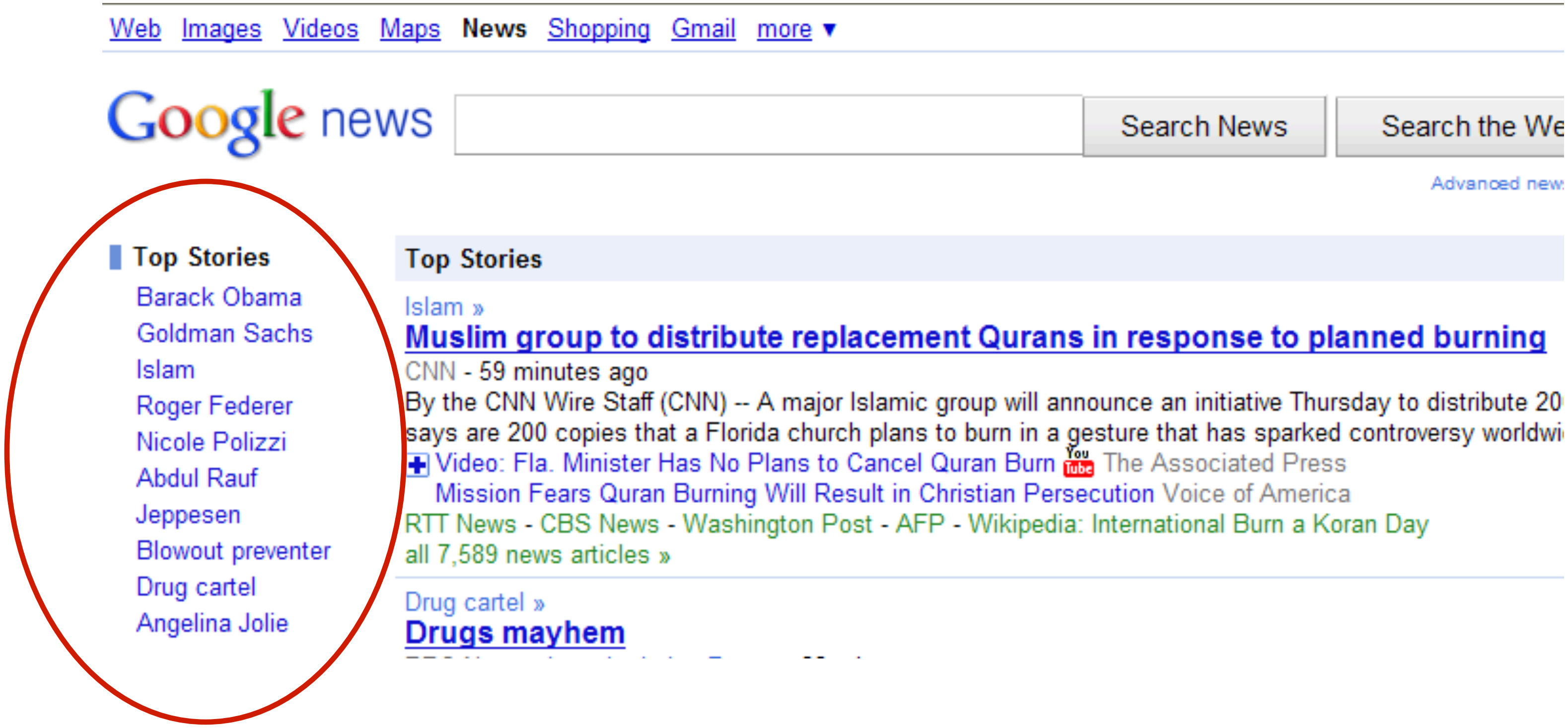
He was slow in learning how to talk. "My parents were so worried," he later recalled, "that they consulted a doctor." Even after he had begun using words, sometime after the age of 2, he developed a quirk that prompted the family maid to dub him "der Depperte," the dopey one, and others in his family to label him as "almost backwards." Whenever he had something to say, he would try it out on himself, whispering it softly until it sounded good enough to pronounce aloud. "Every sentence he uttered," his worshipful younger sister recalled, "no matter how routine, he repeated to himself softly, moving his lips." It was all very worrying, she said. "He had such difficulty with language that those around him feared he would never learn."[1]

His slow development was combined with a cheeky rebelliousness toward authority, which led

3% Locations 412-19 12078

Information Extraction

- Goal: Build database entries from text
- Simple Task: Named Entity Extraction



Top Stories

Barack Obama
Goldman Sachs
Islam
Roger Federer
Nicole Polizzi
Abdul Rauf
Jeppesen
Blowout preventer
Drug cartel
Angelina Jolie

Top Stories

Islam »

Muslim group to distribute replacement Qurans in response to planned burning

CNN - 59 minutes ago

By the CNN Wire Staff (CNN) -- A major Islamic group will announce an initiative Thursday to distribute 20
says are 200 copies that a Florida church plans to burn in a gesture that has sparked controversy worldwi

Video: Fla. Minister Has No Plans to Cancel Quran Burn The Associated Press
Mission Fears Quran Burning Will Result in Christian Persecution Voice of America

RTT News - CBS News - Washington Post - AFP - Wikipedia: International Burn a Koran Day

all 7,589 news articles »

Drug cartel »

Drugs mayhem

Information Extraction

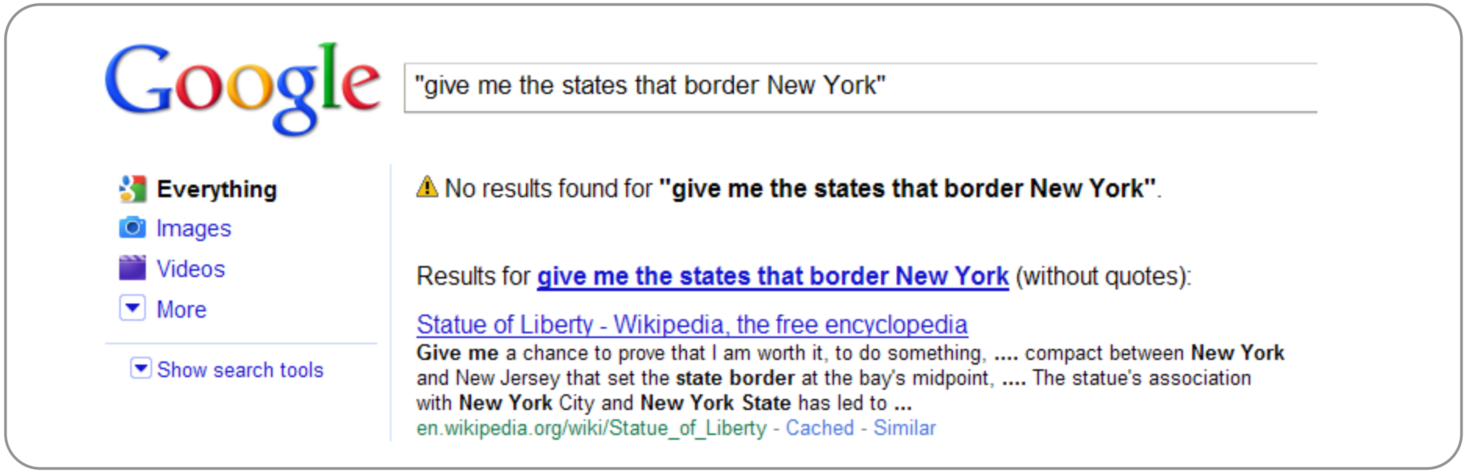
- Goal: Build database entries from text
- More advanced: Multi-sentence Template IE

10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automative account. Experience in online marketing, automative and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables … Compensation: $50,000-\$80,000

INDUSTRY	Advertising
POSITION	Assist. Account Manag.
LOCATION	Irvine, CA
COMPANY	10th DEGREE

Question Answering

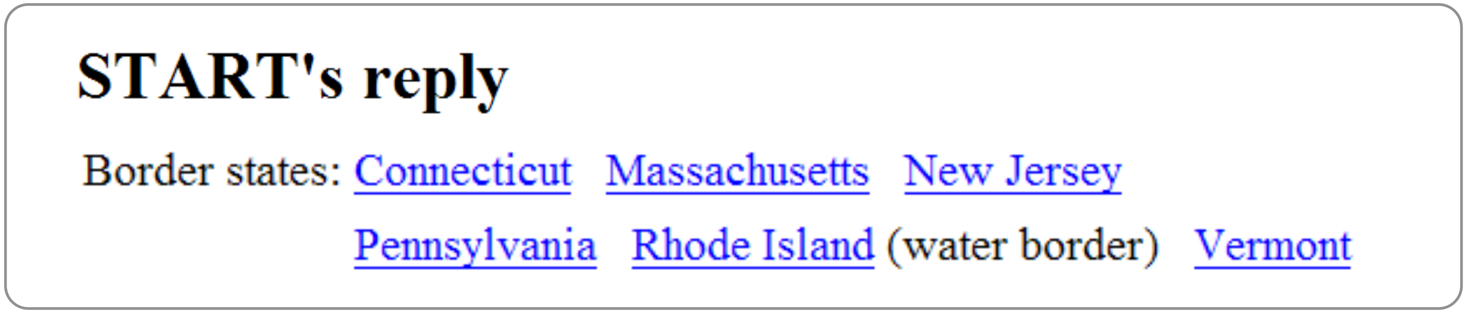
- Find answers to general comprehension question in a document collection



⚠ No results found for **"give me the states that border New York"**.

Results for **give me the states that border New York** (without quotes):

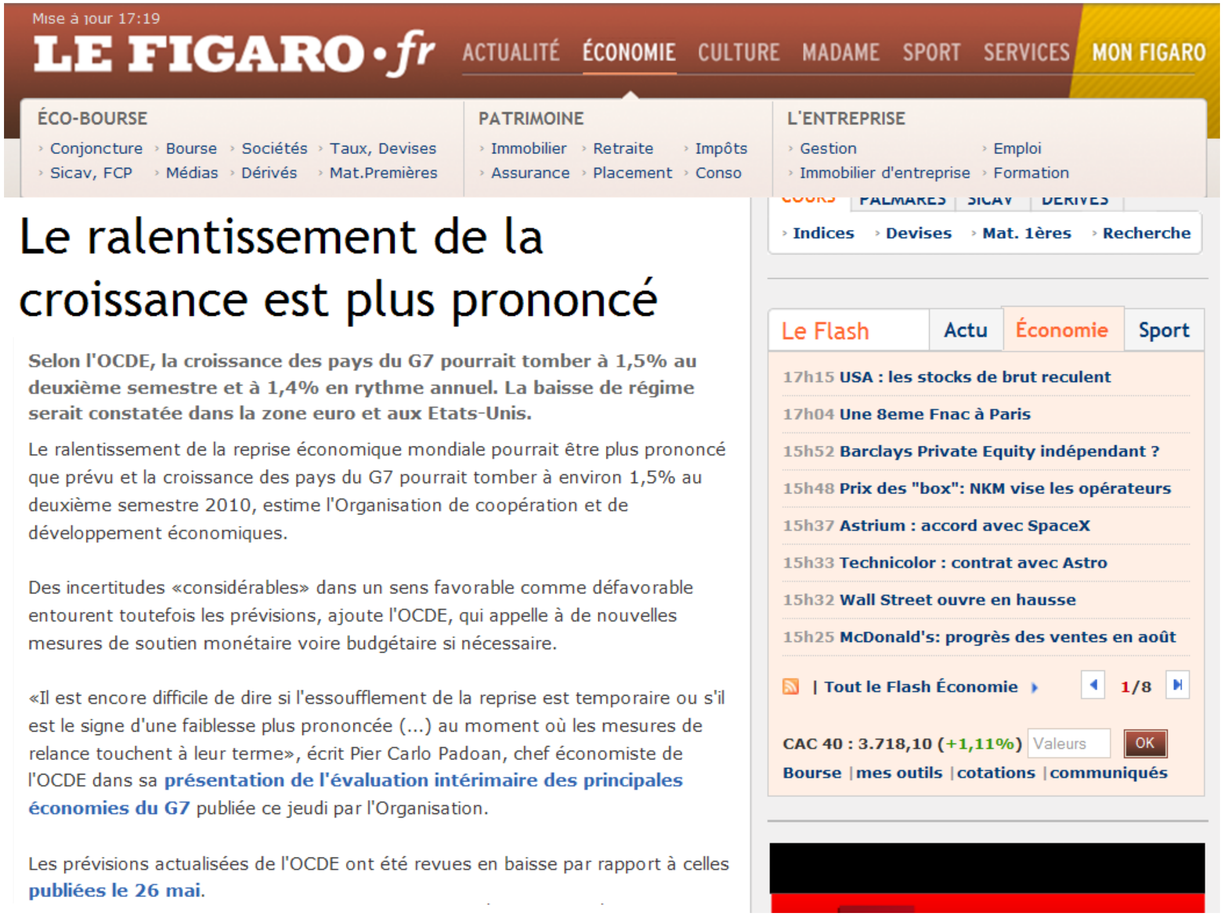
Statue of Liberty - Wikipedia, the free encyclopedia
Give me a chance to prove that I am worth it, to do something, **....** compact between **New York** and New Jersey that set the **state border** at the bay's midpoint, **....** The statue's association with **New York** City and **New York State** has led to **...**
en.wikipedia.org/wiki/Statue_of_Liberty - Cached - Similar



START's reply

Border states: Connecticut Massachusetts New Jersey
Pennsylvania Rhode Island (water border) Vermont

Machine Translation



Le ralentissement de la croissance est plus prononcé

Selon l'OCDE, la croissance des pays du G7 pourrait tomber à 1,5% au deuxième semestre et à 1,4% en rythme annuel. La baisse de régime serait constatée dans la zone euro et aux Etats-Unis.

Le ralentissement de la reprise économique mondiale pourrait être plus prononcé que prévu et la croissance des pays du G7 pourrait tomber à environ 1,5% au deuxième semestre 2010, estime l'Organisation de coopération et de développement économiques.

Des incertitudes «considérables» dans un sens favorable comme défavorable entourent toutefois les prévisions, ajoute l'OCDE, qui appelle à de nouvelles mesures de soutien monétaire voire budgétaire si nécessaire.

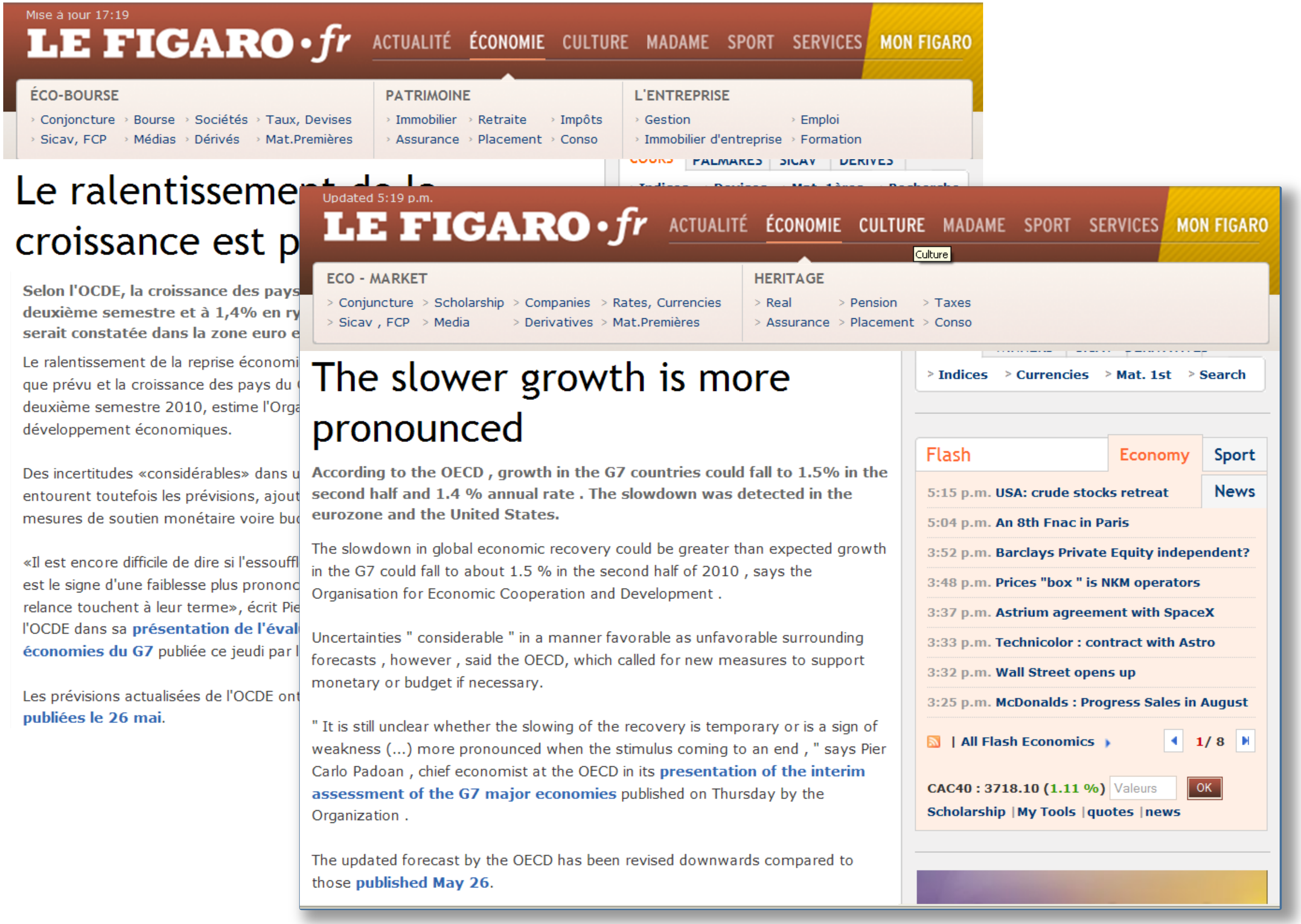
«Il est encore difficile de dire si l'essoufflement de la reprise est temporaire ou s'il est le signe d'une faiblesse plus prononcée (...) au moment où les mesures de relance touchent à leur terme», écrit Pier Carlo Padoan, chef économiste de l'OCDE dans sa **présentation de l'évaluation intérimaire des principales économies du G7** publiée ce jeudi par l'Organisation.

Les prévisions actualisées de l'OCDE ont été revues en baisse par rapport à celles **publiées le 26 mai**.

17h15 **USA : les stocks de brut reculent**
17h04 **Une 8eme Fnac à Paris**
15h52 **Barclays Private Equity indépendant ?**
15h48 **Prix des "box": NKM vise les opérateurs**
15h37 **Astrium : accord avec SpaceX**
15h33 **Technicolor : contrat avec Astro**
15h32 **Wall Street ouvre en hausse**
15h25 **McDonald's: progrès des ventes en août**

CAC 40 : 3.718,10 (+1,11%)

Google Translation



Mise à jour 17:19

LE FIGARO·fr

ACTUALITÉ
ÉCONOMIE
CULTURE
MADAME
SPORT
SERVICES
MON FIGARO
ÉCO-BOURSE
› Conjoncture › Bourse › Sociétés › Taux, Devises
› Sicav, FCP › Médias › Dérivés › Mat.Premières
PATRIMOINE
› Immobilier › Retraite › Impôts
› Assurance › Placement › Conso
L'ENTREPRISE
› Gestion › Emploi
› Immobilier d'entreprise › Formation
COURS PALMARÈS SICAV DÉRIVÉS

Le ralentissement de la croissance est p

Selon l'OCDE, la croissance des pays deuxième semestre et à 1,4% en ry serait constatée dans la zone euro e

Le ralentissement de la reprise économi que prévu et la croissance des pays du deuxième semestre 2010, estime l'Orga développement économiques.

Des incertitudes «considérables» dans u entourent toutefois les prévisions, ajout mesures de soutien monétaire voire bu

«Il est encore difficile de dire si l'essouffl est le signe d'une faiblesse plus pronon relance touchent à leur terme», écrit Pie l'OCDE dans sa **présentation de l'éval économies du G7** publiée ce jeudi par l

Les prévisions actualisées de l'OCDE ont **publiées le 26 mai**.

Updated 5:19 p.m.

LE FIGARO·fr

ACTUALITÉ
ÉCONOMIE
CULTURE
MADAME
SPORT
SERVICES
MON FIGARO
Culture
ECO - MARKET
> Conjuncture > Scholarship > Companies > Rates, Currencies
> Sicav , FCP > Media > Derivatives > Mat.Premières
HERITAGE
> Real > Pension > Taxes
> Assurance > Placement > Conso
> Indices > Currencies > Mat. 1st > Search

The slower growth is more pronounced

According to the OECD , growth in the G7 countries could fall to 1.5% in the second half and 1.4 % annual rate . The slowdown was detected in the eurozone and the United States.

The slowdown in global economic recovery could be greater than expected growth in the G7 could fall to about 1.5 % in the second half of 2010 , says the Organisation for Economic Cooperation and Development .

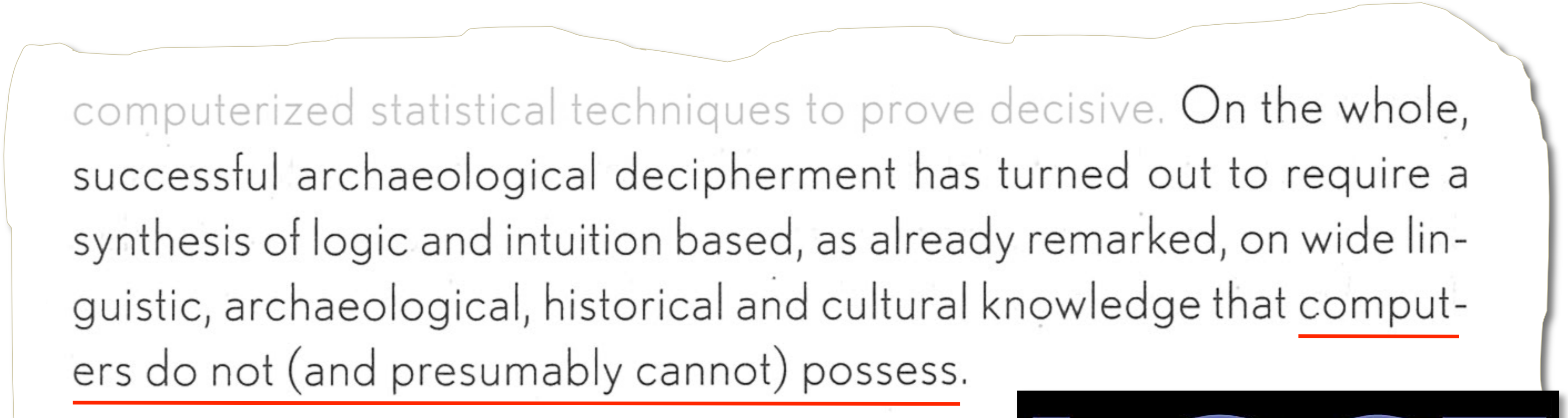
Uncertainties " considerable " in a manner favorable as unfavorable surrounding forecasts , however , said the OECD, which called for new measures to support monetary or budget if necessary.

" It is still unclear whether the slowing of the recovery is temporary or is a sign of weakness (...) more pronounced when the stimulus coming to an end , " says Pier Carlo Padoan , chief economist at the OECD in its **presentation of the interim assessment of the G7 major economies** published on Thursday by the Organization .

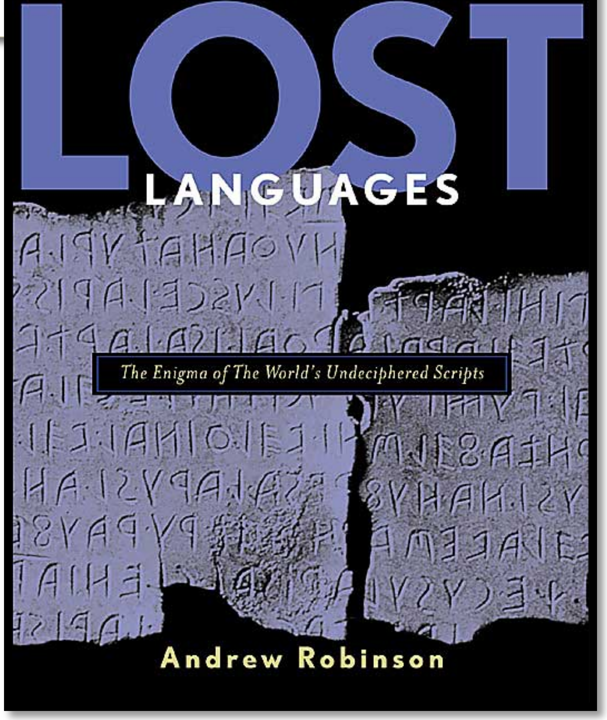
The updated forecast by the OECD has been revised downwards compared to those **published May 26**.

5:15 p.m. **USA: crude stocks retreat**
5:04 p.m. **An 8th Fnac in Paris**
3:52 p.m. **Barclays Private Equity independent?**
3:48 p.m. **Prices "box " is NKM operators**
3:37 p.m. **Astrium agreement with SpaceX**
3:33 p.m. **Technicolor : contract with Astro**
3:32 p.m. **Wall Street opens up**
3:25 p.m. **McDonalds : Progress Sales in August**

CAC40 : 3718.10 (1.11 %)



computerized statistical techniques to prove decisive. On the whole, successful archaeological decipherment has turned out to require a synthesis of logic and intuition based, as already remarked, on wide linguistic, archaeological, historical and cultural knowledge that computers do not (and presumably cannot) possess.

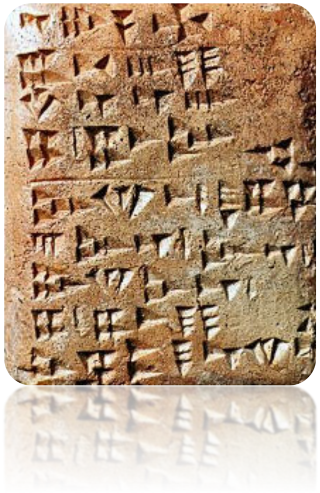


LOST LANGUAGES

The Enigma of The World's Undeciphered Scripts

Andrew Robinson

Deciphering Ugaritic



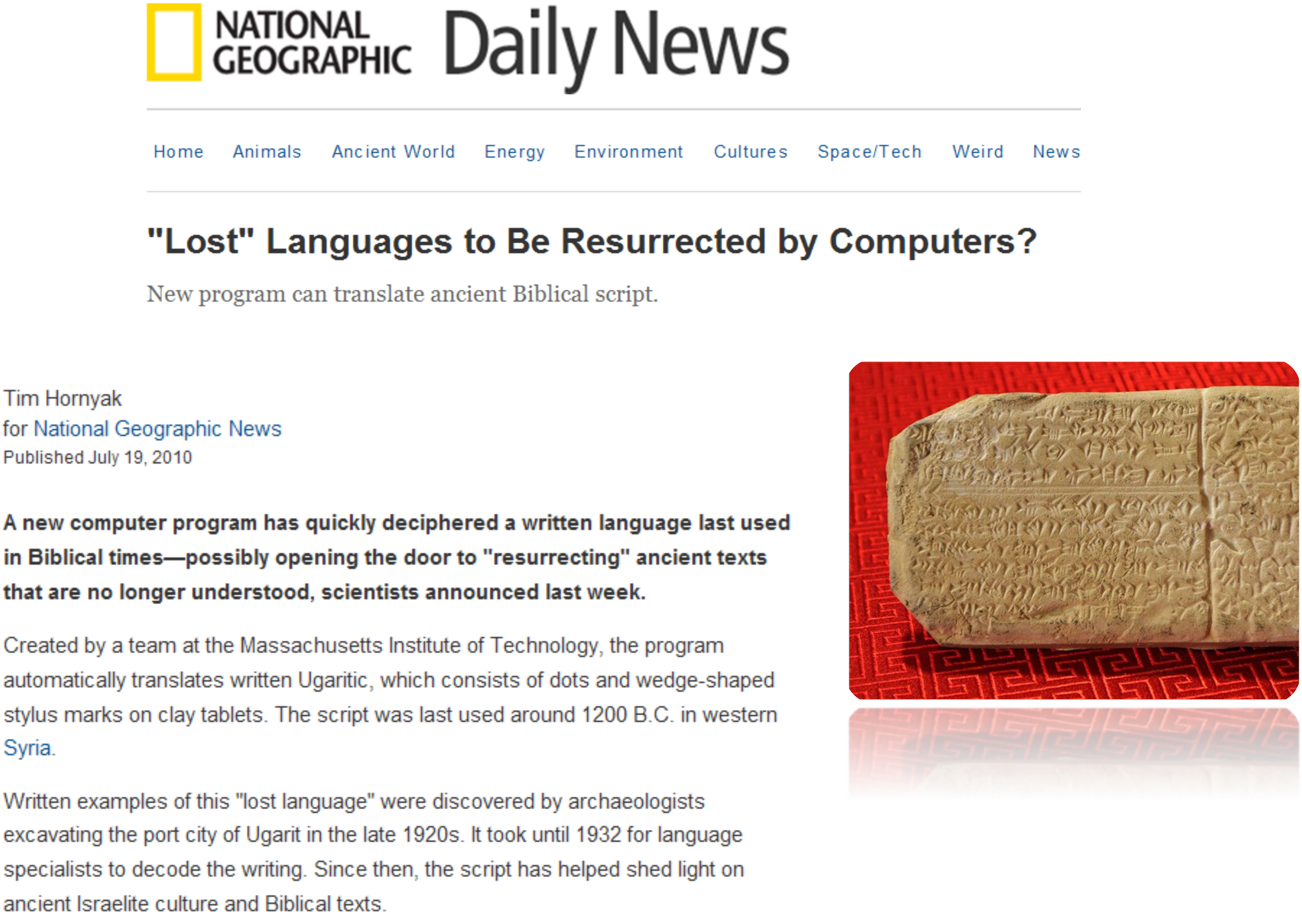
Family	: Northwest Semitic
Tablets from	: 14th – 12th century BCE
Discovered	: 1928
Deciphered	: 1932 (by WW1 code breakers)

Large portion of vocabulary covered by cognates with Semitic languages

Arabic:	malik	مَلك
Syriac:	malkā	
		𐎎𐎍𐎋
Hebrew:	melek	מֶלֶךְ
Ugaritic:	malku	

Task: Translate by identifying cognates

Corpus: 34,105 tokens, 7,386 unique types



"Lost" Languages to Be Resurrected by Computers?

New program can translate ancient Biblical script.

Tim Hornyak
for National Geographic News
Published July 19, 2010

A new computer program has quickly deciphered a written language last used in Biblical times—possibly opening the door to "resurrecting" ancient texts that are no longer understood, scientists announced last week.

Created by a team at the Massachusetts Institute of Technology, the program automatically translates written Ugaritic, which consists of dots and wedge-shaped stylus marks on clay tablets. The script was last used around 1200 B.C. in western Syria.

Written examples of this "lost language" were discovered by archaeologists excavating the port city of Ugarit in the late 1920s. It took until 1932 for language specialists to decode the writing. Since then, the script has helped shed light on ancient Israelite culture and Biblical texts.

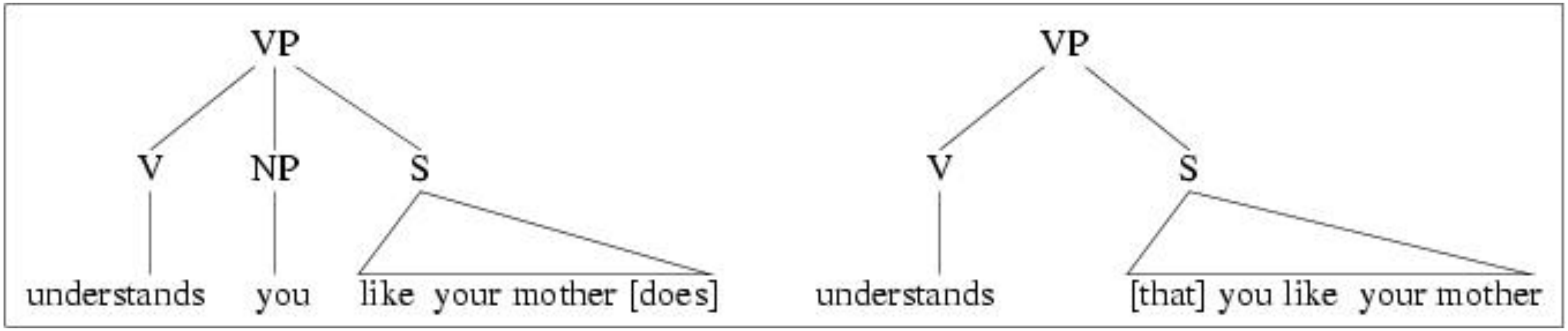
Why are these funny?

- Iraqi Head Seeks Arms
- Ban on Nude Dancing on Governor's Desk
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snaks
- Local HS Dropout Cut in Half
- Hospitals Are Sued by 7 Foot Doctors

Why NLP is Hard?
(example from L.Lee)

``At last, a computer that understands you like your mother''

Ambiguity at Syntactic Level



Different structures lead to different interpretations

Ambiguity at Semantic Level

“Alice says they've built a computer that understands you like your mother”

Two definitions of mother:

- female parent
- a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

This is an instance of word sense disambiguation

Ambiguity at Discourse Level

Alice says they've built a computer that understands you like your mother but she

- … doesn’t know any details
- … doesn’t understand me at all

This is an instance of anaphora, where “she” co-refers to some other discourse entity

Ambiguity Varies Across Languages

- Tokenization

English: in the country

Hebrew: ב

Easy task in English: space separator delineates words.

Challenging for Semitic Languages

- Named Entity Detection

English: She saw Jacob …

Hebrew: ה

Easy task in English: capitalization is a strong hint.

Challenging for Semitic languages.

Knowledge Bottleneck in NLP

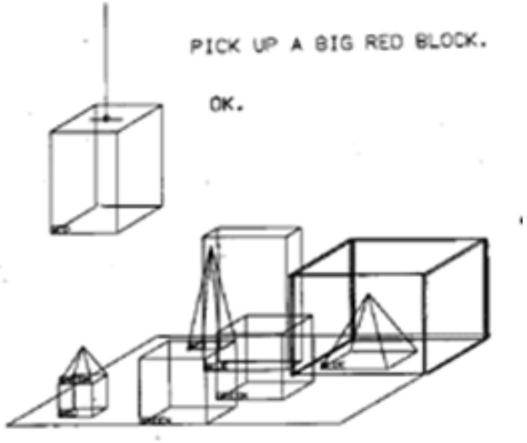
We need:

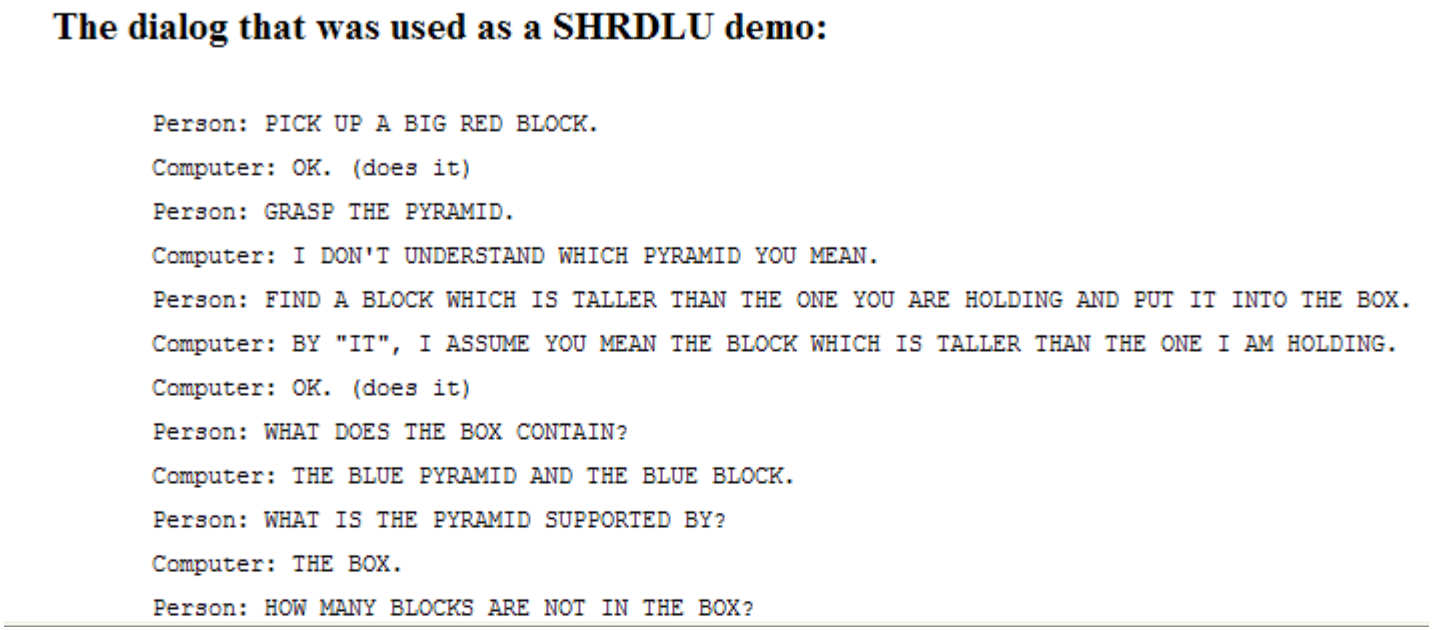
- Knowledge about language
- Knowledge about the world

Possible solutions:

- Symbolic approach: encode all the required information into computer
- Statistical approach: infer language properties from language samples

Symbolic Era: Crowning Achievement



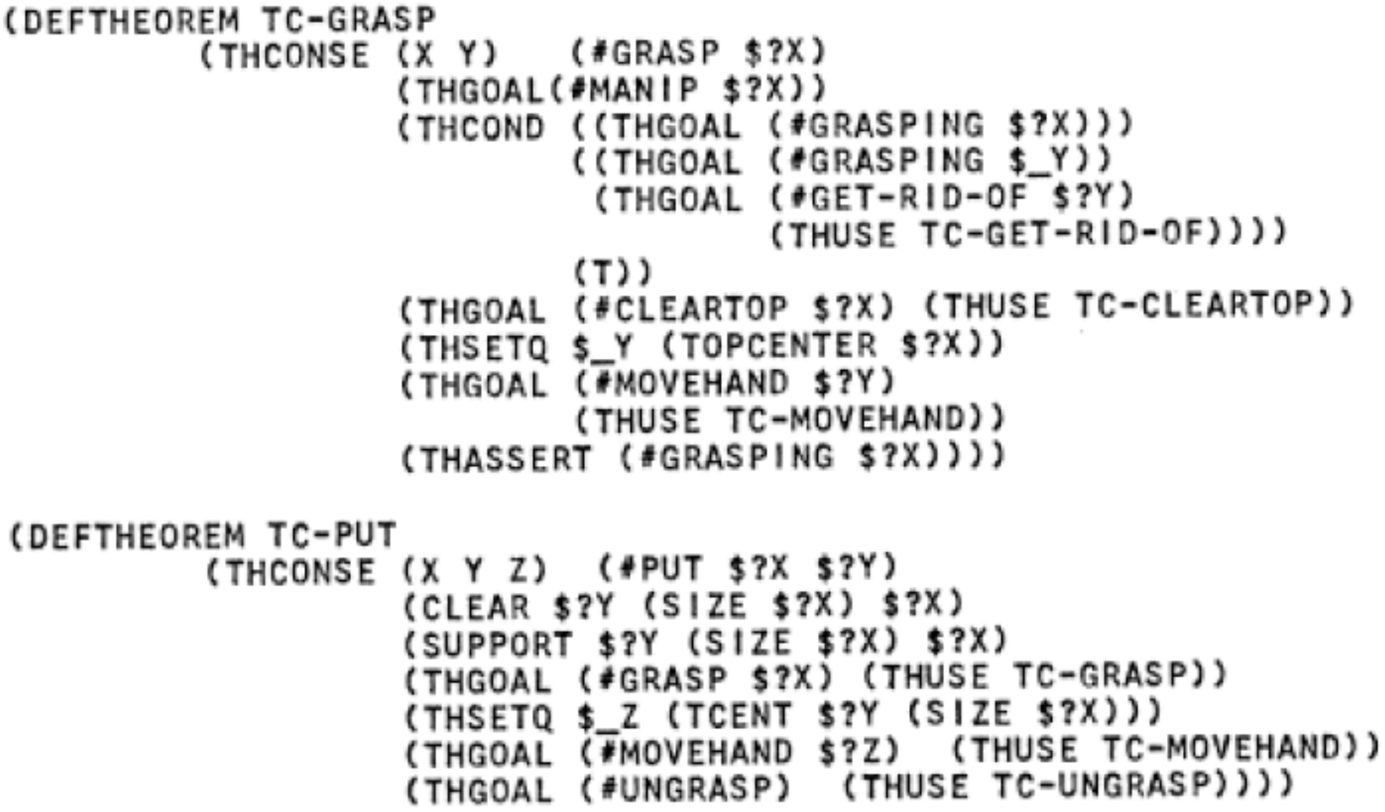


The dialog that was used as a SHRDLU demo:

```
Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
```

The Internals of *SHRDLU*

Requires elaborate manually encoded knowledge representation



```
(DEFTHEOREM TC-GRASP
         (THCONSE (X Y)   (#GRASP $?X)
                  (THGOAL(#MANIP $?X))
                  (THCOND ((THGOAL (#GRASPING $?X)))
                          ((THGOAL (#GRASPING $_Y))
                           (THGOAL (#GET-RID-OF $?Y)
                                   (THUSE TC-GET-RID-OF))))
                          (T))
                  (THGOAL (#CLEARTOP $?X) (THUSE TC-CLEARTOP))
                  (THSETQ $_Y (TOPCENTER $?X))
                  (THGOAL (#MOVEHAND $?Y)
                          (THUSE TC-MOVEHAND))
                  (THASSERT (#GRASPING $?X))))

(DEFTHEOREM TC-PUT
         (THCONSE (X Y Z)  (#PUT $?X $?Y)
                  (CLEAR $?Y (SIZE $?X) $?X)
                  (SUPPORT $?Y (SIZE $?X) $?X)
                  (THGOAL (#GRASP $?X) (THUSE TC-GRASP))
                  (THSETQ $_Z (TCENT $?Y (SIZE $?X)))
                  (THGOAL (#MOVEHAND $?Z)  (THUSE TC-MOVEHAND))
                  (THGOAL (#UNGRASP)  (THUSE TC-UNGRASP))))
```

NLP History: Symbolic Era

> *"Colorless green ideas sleep furiously.*
>
> *Furiously sleep ideas green colorless.*
>
> *It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not."* (Chomsky 1957)

1970's and 1980's: statistical NLP is in disfavor

- emphasis on deeper models, syntax
- toy domains/manually developed grammars (SHRDLU, LUNAR)
- weak empirical evaluation

NLP History: Statistical Era

"*Whenever I fire a linguist our system performance improves.* " (Jelinek 1988)

1990's: The Empirical Revolution

- Corpus-based methods yield the first generation of NL tools (syntax, MT, ASR)
- Deep analysis is often traded for robust approximations
- Empirical evaluation is crucial

2000's: Richer linguistic representations embedded in the statistical framework

Case Study: Determiner Placement

Task: Automatically place determiners *a, the, null* in a text

Scientists in United States have found way of turning lazy monkeys into workaholics using gene therapy. Usually monkeys work hard only when they know reward is coming, but animals given this treatment did their best all time. Researchers at National Institute of Mental Health near Washington DC, led by Dr Barry Richmond, have now developed genetic treatment which changes their work ethic markedly. "Monkeys under influence of treatment don't procrastinate," Dr Richmond says. Treatment consists of anti-sense DNA - mirror image of piece of one of our genes - and basically prevents that gene from working. But for rest of us, day when such treatments fall into hands of our bosses may be one we would prefer to put off.

Relevant Grammar Rules

- Determiner placement is largely determined by:
  - Type of noun (countable, uncountable)
  - Uniqueness of reference
  - Information value (given, new)
  - Number (singular, plural)
- However, many exceptions and special cases play a role:
  - The definite article is used with newspaper titles (The Times), but zero article in names of magazines and journals (Time)

Hard to manually encode this information!

Statistical Approach: Determiner Placement

Simple approach:

- Collect a large collection of texts relevant to your domain (e.g. newspaper text)
- For each noun seen during training, compute its probability to take a certain determiner
- Given a new noun, select a determiner with the highest likelihood as estimated on the training corpus

Determiner Placement as Classification

- **Prediction**: *{``the'', ``a'', ``null''}*
- **Representation of the problem**:
  - plural? (yes, no)
  - first appearance in text? (yes, no)
  - head token (vocabulary)

Plural?	First appearance?	Token	Determiner
no	yes	defendant	the
yes	no	cars	null
no	no	FBI	the

Goal: Learn classification function that can predict unseen examples

Does it work?

- Implementation details:
  - Training --- first 21 sections of the Wall Street Journal corpus, testing -- the 23th section
  - Prediction accuracy: 71.5%
- The results are not great, but surprisingly high for such a simple method
  - A large fraction of nouns in this corpus always appear with the same determiner

``the FBI'', ``the defendant''

Corpora

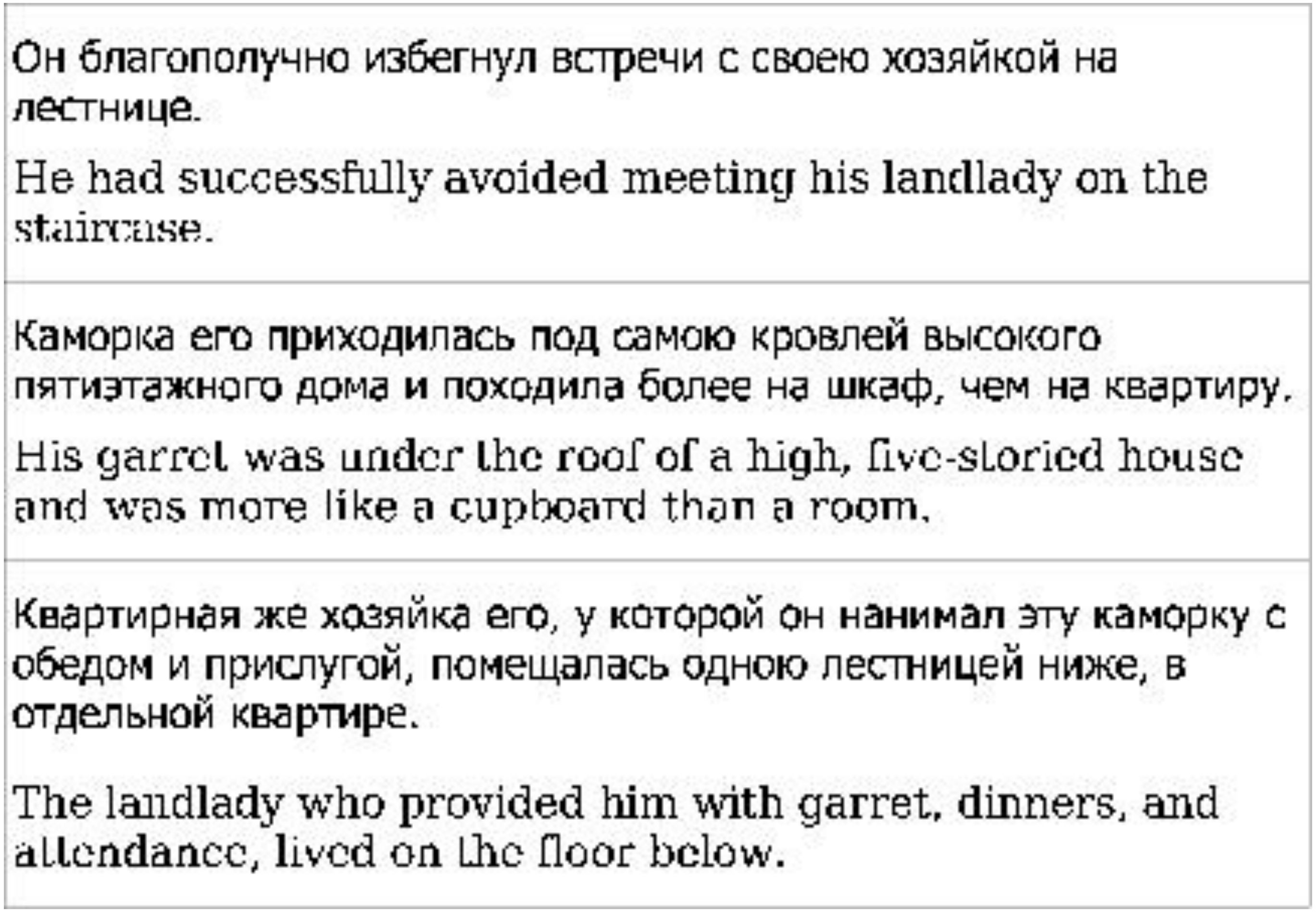
Corpus: a collection of annotated or raw text

Antique corpus: Rosetta Stone

Examples of corpora used in NLP today:

- Penn Treebank: 1M words of parsed text
- Brown Corpus: 1M words of tagged text
- North American News: 300M words
- The Web

Corpus for MT



Он благополучно избегнул встречи с своею хозяйкой на лестнице.

He had successfully avoided meeting his landlady on the staircase.

Каморка его приходилась под самою кровлей высокого пятиэтажного дома и походила более на шкаф, чем на квартиру.

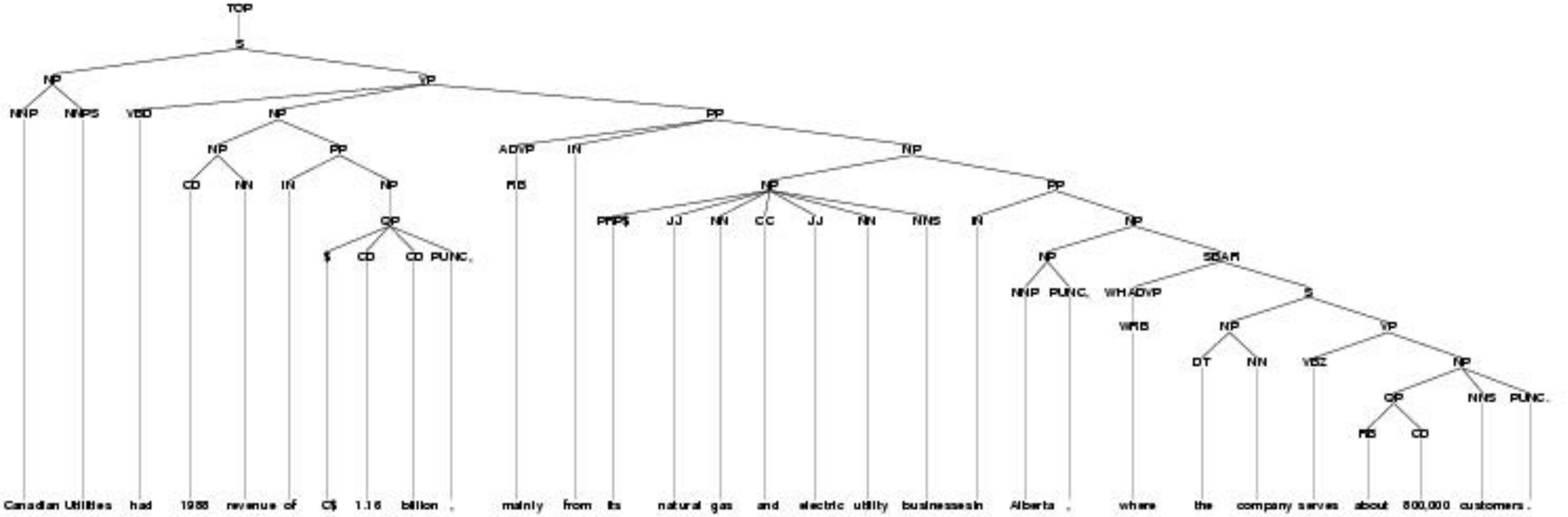
His garret was under the roof of a high, five-storied house and was more like a cupboard than a room.

Квартирная же хозяйка его, у которой он нанимал эту каморку с обедом и прислугой, помещалась одною лестницей ниже, в отдельной квартире.

The landlady who provided him with garret, dinners, and attendance, lived on the floor below.

Corpus for Parsing

Canadian Utilities had 1988 revenue of $ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers **.**

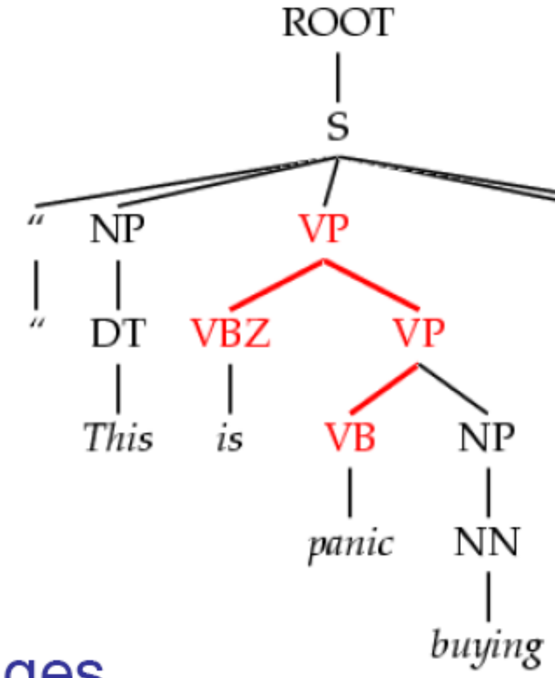


Ambiguities

- *Dark ambiguities*: most analyses are shockingly bad (meaning, they don't have an interpretation you can get your mind around)

This analysis corresponds to the correct parse of

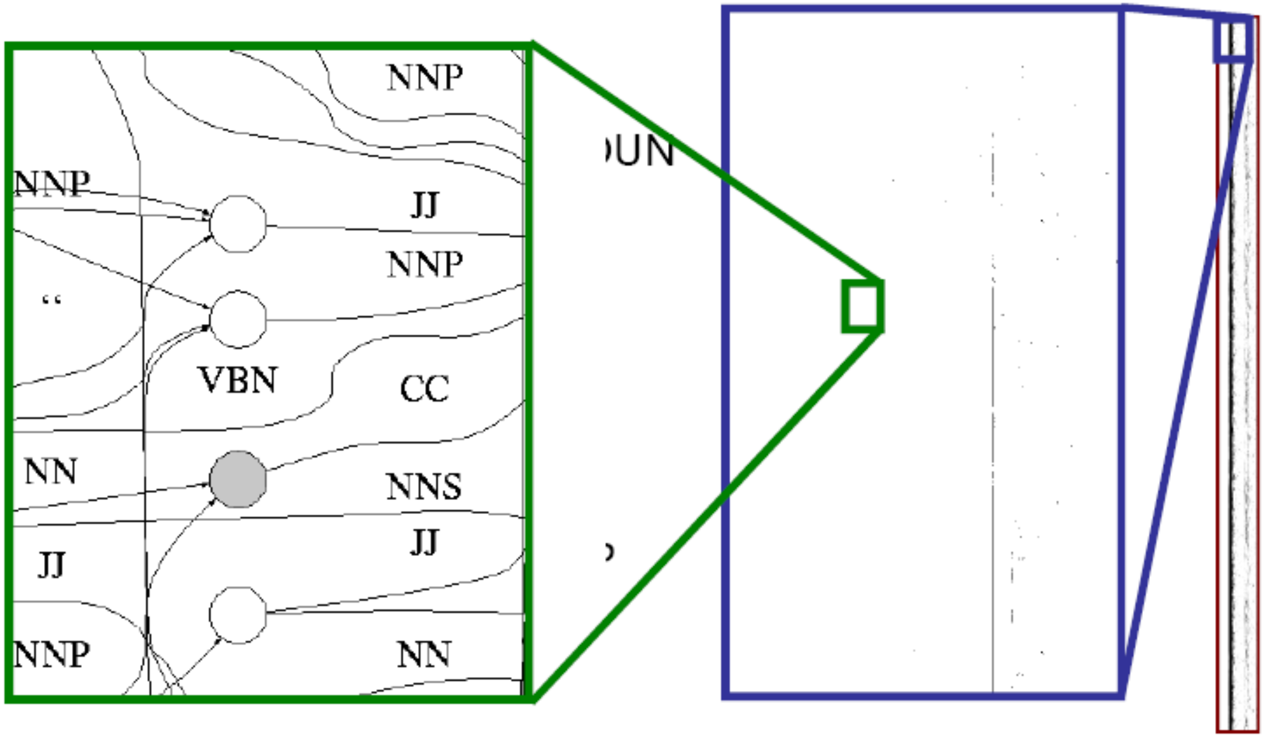
"This will panic buyers ! "



- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

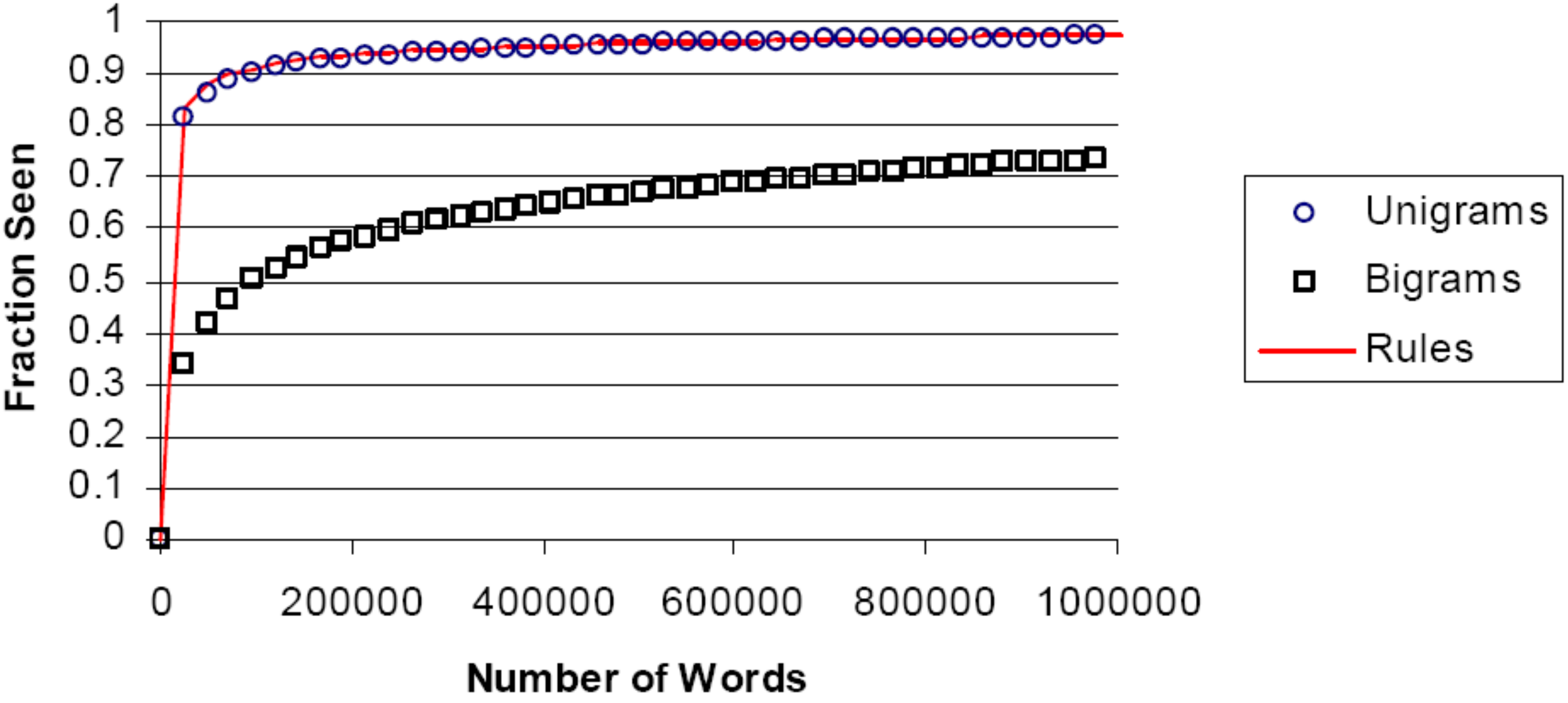
Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - …they didn’t realize how bad it would be



Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire



The NLP Cycle

- Get a corpus
- Build a baseline model
- Repeat:
  - Analyze the most common errors
  - Find out what information could be helpful
  - Modify the model to exploit this information
    - Use new features
    - Change the structure of the model
    - Employ new machine learning method

Parsing and Syntax

Syntactic Formalisms: Historic Perspective

- "Syntax" comes from Greek word "*syntaxis*", meaning "setting out together or arrangement"
- Early grammars: 4th century BC
  - Panini compiled Sanskrit grammar
- Idea of constituency
  - Bloomfield (1914): method for breaking up sentence into a hierarchy of units
  - Harris (1954): substitutability test for constituent definition
- Formal work on Syntax goes back to Chomsky's PhD thesis in 1950s

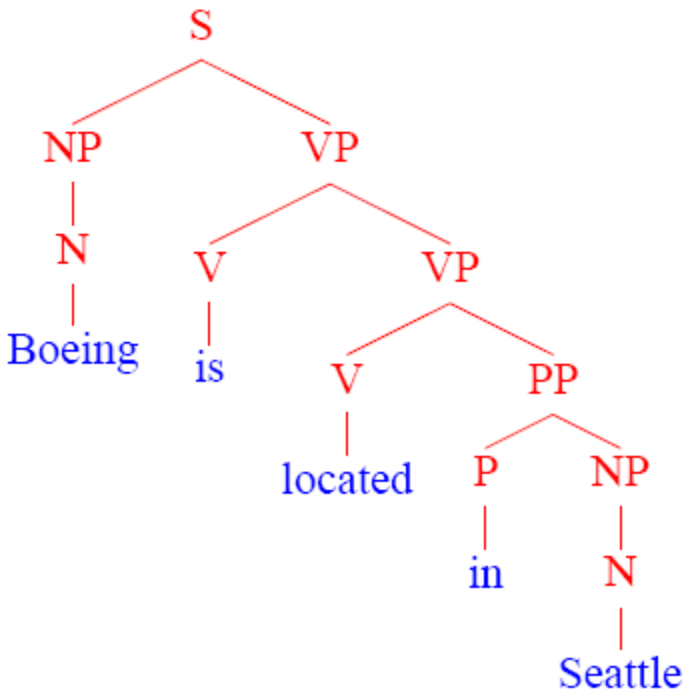
*Some slides in this lecture are adapted from slides of Michael Collins

Syntactic Structure

INPUT:

Boeing is located in Seattle.

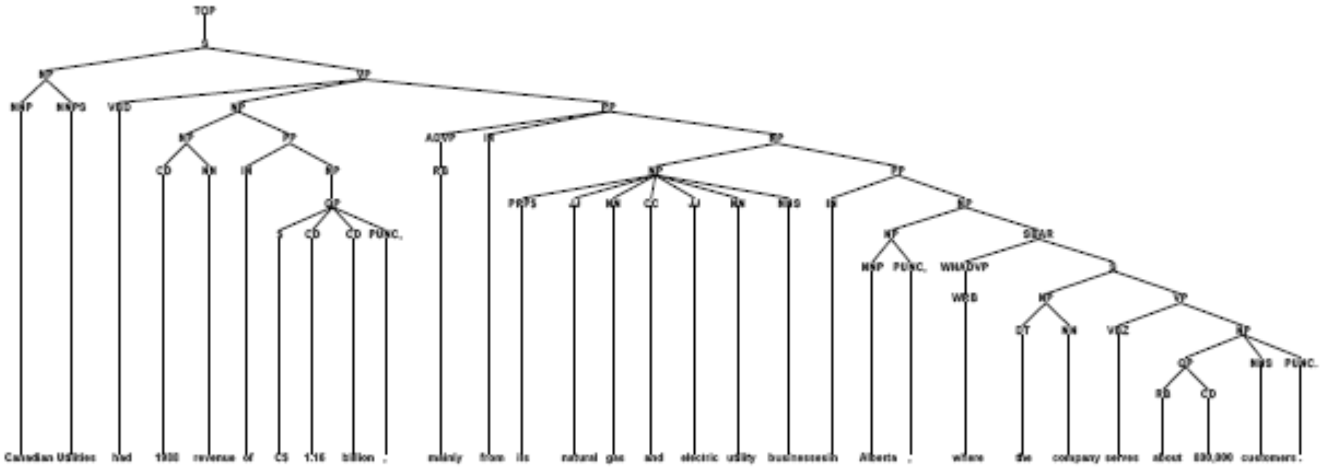
OUTPUT:



A Real Tree

- Penn WSJ Treebank = 50,000 sentences with associated trees
- Usual set-up: 40,000 training sentences, 2400 test sentences

An example tree:



Canadian Utilities had 1988 revenue of C$ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers .

What can we Learn from Syntactic Tree?

- Part-of-speech for each word

  (N=noun, V=verb, P=preposition)

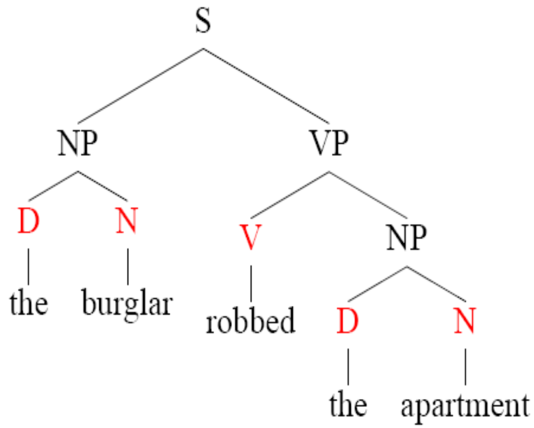
- Constituent structure

  Noun phrase: “the apartment”

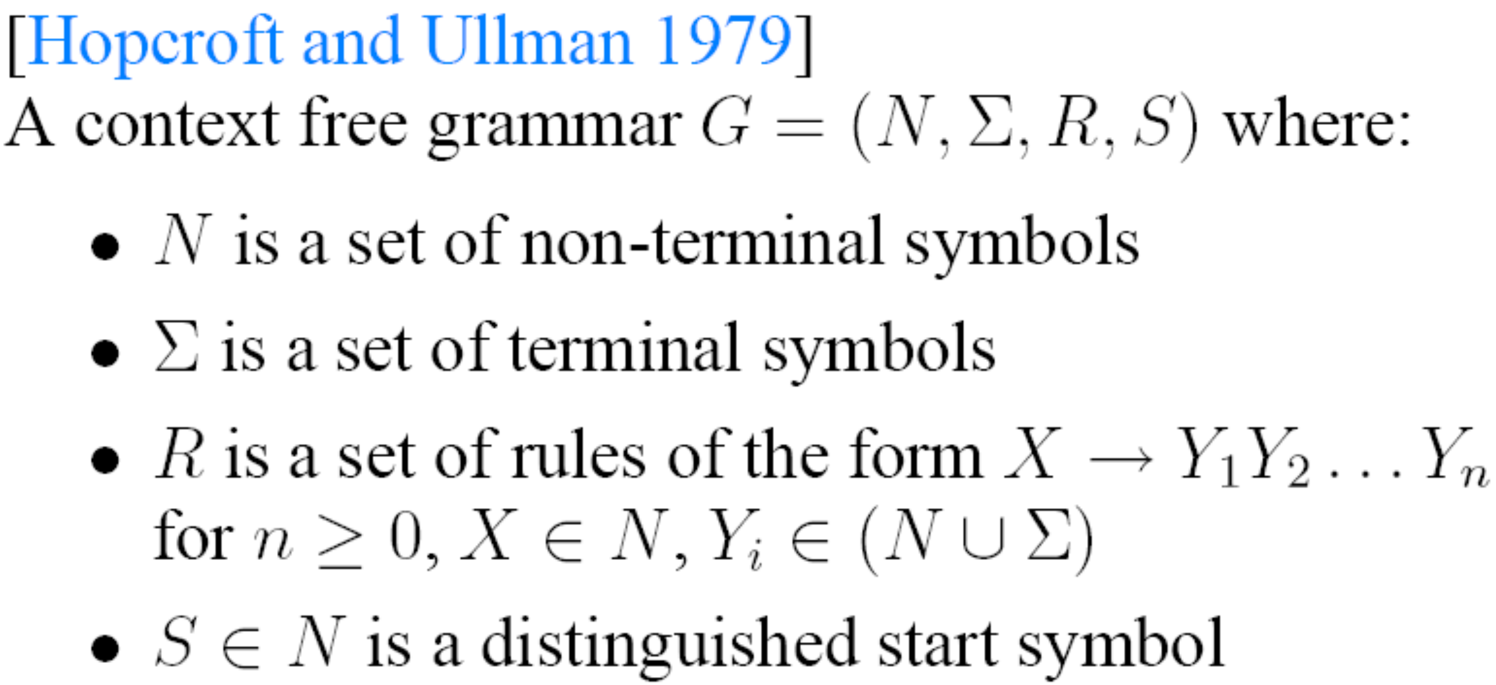
  Verb phrase: “robbed the apartment”

- Relationship structure

  “the burglar” is the subject of “robbed”



Context-Free Grammars



[Hopcroft and Ullman 1979]
A context free grammar $G = (N, \Sigma, R, S)$ where:

- N is a set of non-terminal symbols
- Σ is a set of terminal symbols
- R is a set of rules of the form $X \rightarrow Y_1 Y_2 \ldots Y_n$ for $n \geq 0$, $X \in N$, $Y_i \in (N \cup \Sigma)$
- $S \in N$ is a distinguished start symbol

A Context-Free Grammar for English

$N = \{\text{S, NP, VP, PP, DT, Vi, Vt, NN, IN}\}$

$S = \text{S}$

$\Sigma = \{\text{sleeps, saw, man, woman, telescope, the, with, in}\}$

$R =$

S	⇒	NP	VP
VP	⇒	Vi	
VP	⇒	Vt	NP
VP	⇒	VP	PP
NP	⇒	DT	NN
NP	⇒	NP	PP
PP	⇒	IN	NP

Vi	⇒	sleeps
Vt	⇒	saw
NN	⇒	man
NN	⇒	woman
NN	⇒	telescope
DT	⇒	the
IN	⇒	with
IN	⇒	in

Note: S=sentence, VP=verb phrase, NP=noun phrase, PP=prepositional phrase, DT=determiner, Vi=intransitive verb, Vt=transitive verb, NN=noun, IN=preposition

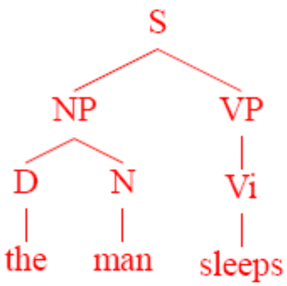
Left-Most Derivation

A left-most derivation is a sequence of strings $s_1 \ldots s_n$, where

- $s_1 = S$, the start symbol
- $s_n \in \Sigma^*$, i.e. s_n is made up of terminal symbols only
- Each s_i for $i = 2 \ldots n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \rightarrow \beta$ is a rule in R

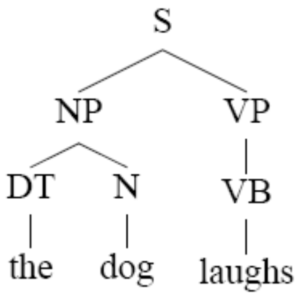
For example: [S], [NP VP], [D N VP], [the N VP], [the man VP], [the man Vi], [the man sleeps]

Representation of a derivation as a tree:

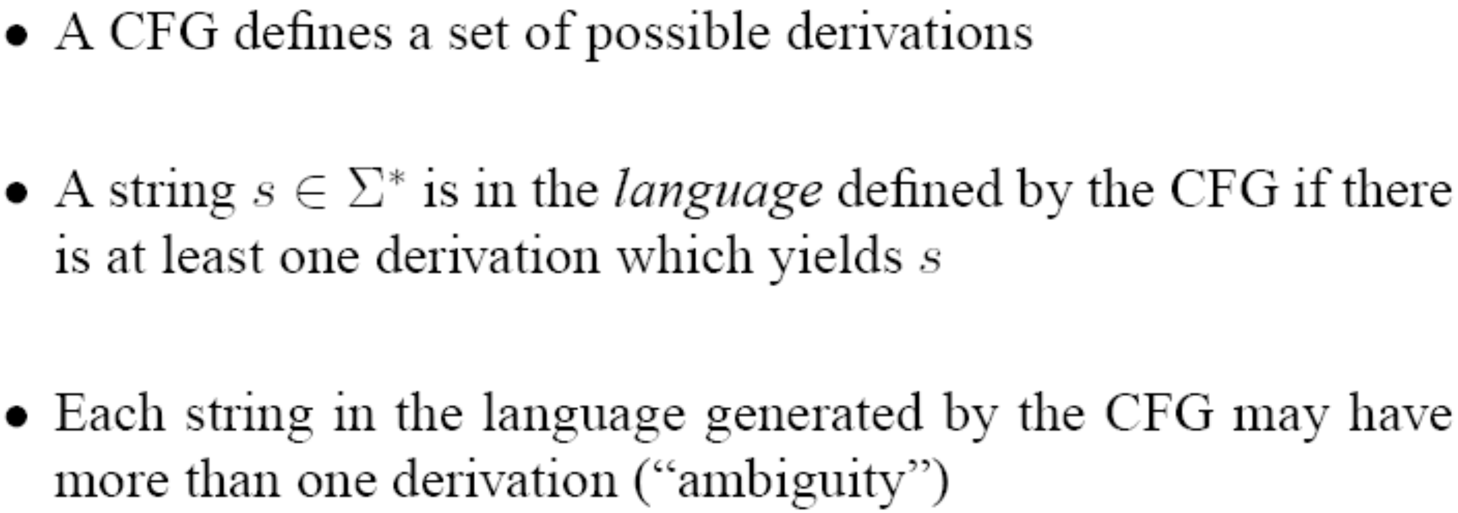


Derivation Example

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → DT N
DT N VP	DT → the
the N VP	N → dog
the dog VP	VP → VB
the dog VB	VB → laughs
the dog laughs	



Properties of CFGs



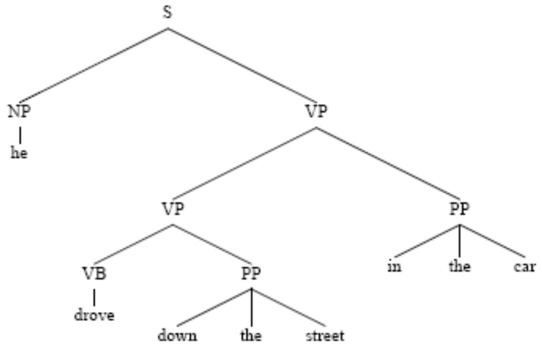
- A CFG defines a set of possible derivations

- A string $s \in \Sigma^*$ is in the *language* defined by the CFG if there is at least one derivation which yields s

- Each string in the language generated by the CFG may have more than one derivation (“ambiguity”)

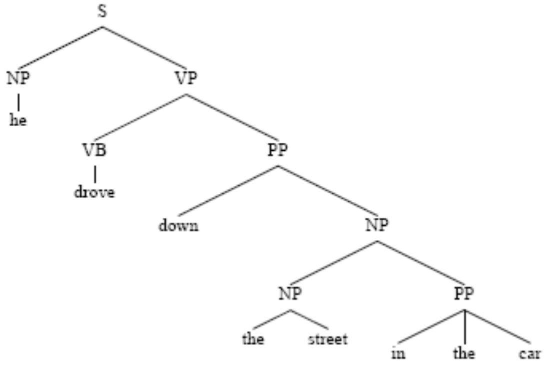
Ambiguous Sentence

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → he
he VP	VP → VP PP
he VP PP	VP → VB PP
he VB PP PP	VB→ drove
he drove PP PP	PP→ down the street
he drove down the street PP	PP→ in the car
he drove down the street in the car	



Ambiguous Sentence

DERIVATION	RULES USED
S	S → NP VP
NP VP	NP → he
he VP	VP → VB PP
he VB PP	VB → drove
he drove PP	PP → down NP
he drove down NP	NP → NP PP
he drove down NP PP	NP → the street
he drove down the street PP	PP → in the car
he drove down the street in the car	

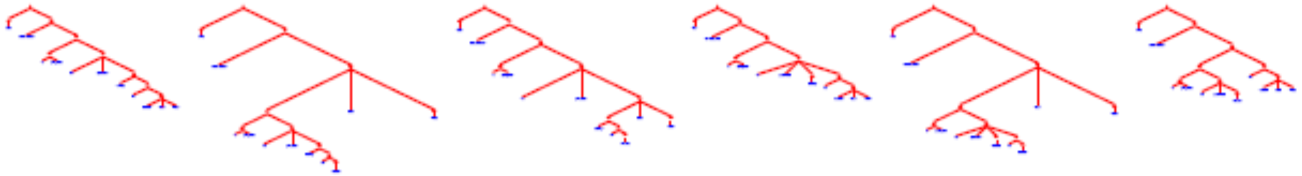


More Ambiguity

She announced a program to promote safety in trucks and vans

⇓

POSSIBLE OUTPUTS:



And there are more...

Syntactic Ambiguity

- Prepositional phrases

  They cooked the beans in the pot on the stove with handles.
- Particle vs preposition

  The puppy tore up the staircase.
- Complement structure

  She knows you like the back of her hand.
- Gerund vs. participial adjective.

  Visiting relatives can be boring
- Modifier scope within NPs

  Plastic cup holder

(examples are compiled by Dan Klein)

Human Processing

- Garden Path:

  The horse raced past the barn fell.

  The man who hunts ducks out on weekends

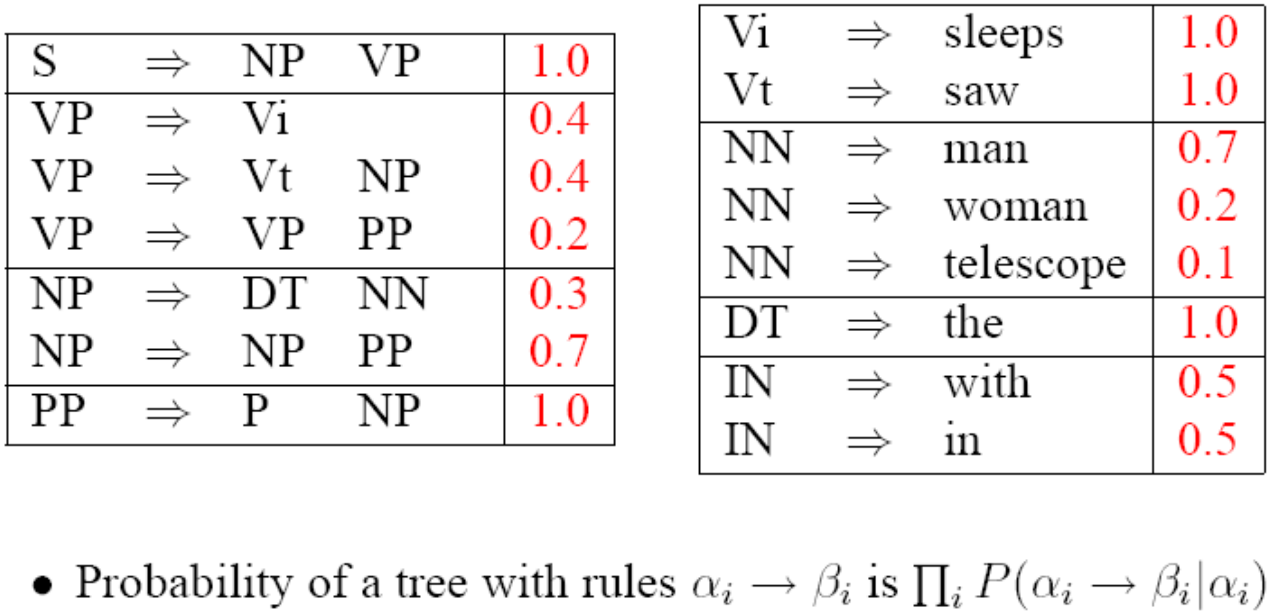
- Ambiguity maintenance

  Have the police … eaten their supper?

  come in and look around

  taken out and shot

A Probabilistic Context-Free Grammar

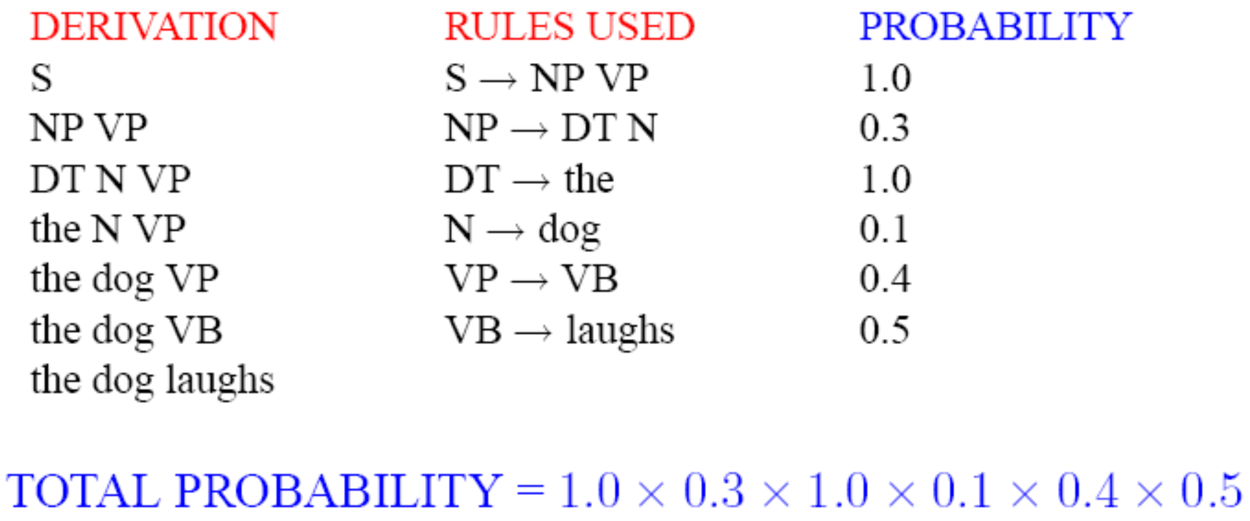


S	$\Rightarrow$	NP	VP	1.0
VP	$\Rightarrow$	Vi		0.4
VP	$\Rightarrow$	Vt	NP	0.4
VP	$\Rightarrow$	VP	PP	0.2
NP	$\Rightarrow$	DT	NN	0.3
NP	$\Rightarrow$	NP	PP	0.7
PP	$\Rightarrow$	P	NP	1.0

Vi	$\Rightarrow$	sleeps	1.0
Vt	$\Rightarrow$	saw	1.0
NN	$\Rightarrow$	man	0.7
NN	$\Rightarrow$	woman	0.2
NN	$\Rightarrow$	telescope	0.1
DT	$\Rightarrow$	the	1.0
IN	$\Rightarrow$	with	0.5
IN	$\Rightarrow$	in	0.5

- Probability of a tree with rules $\alpha_i \rightarrow \beta_i$ is $\prod_i P(\alpha_i \rightarrow \beta_i | \alpha_i)$

Example



DERIVATION	RULES USED	PROBABILITY
S	S → NP VP	1.0
NP VP	NP → DT N	0.3
DT N VP	DT → the	1.0
the N VP	N → dog	0.1
the dog VP	VP → VB	0.4
the dog VB	VB → laughs	0.5
the dog laughs		

TOTAL PROBABILITY $= 1.0 \times 0.3 \times 1.0 \times 0.1 \times 0.4 \times 0.5$

Properties of PCFGs

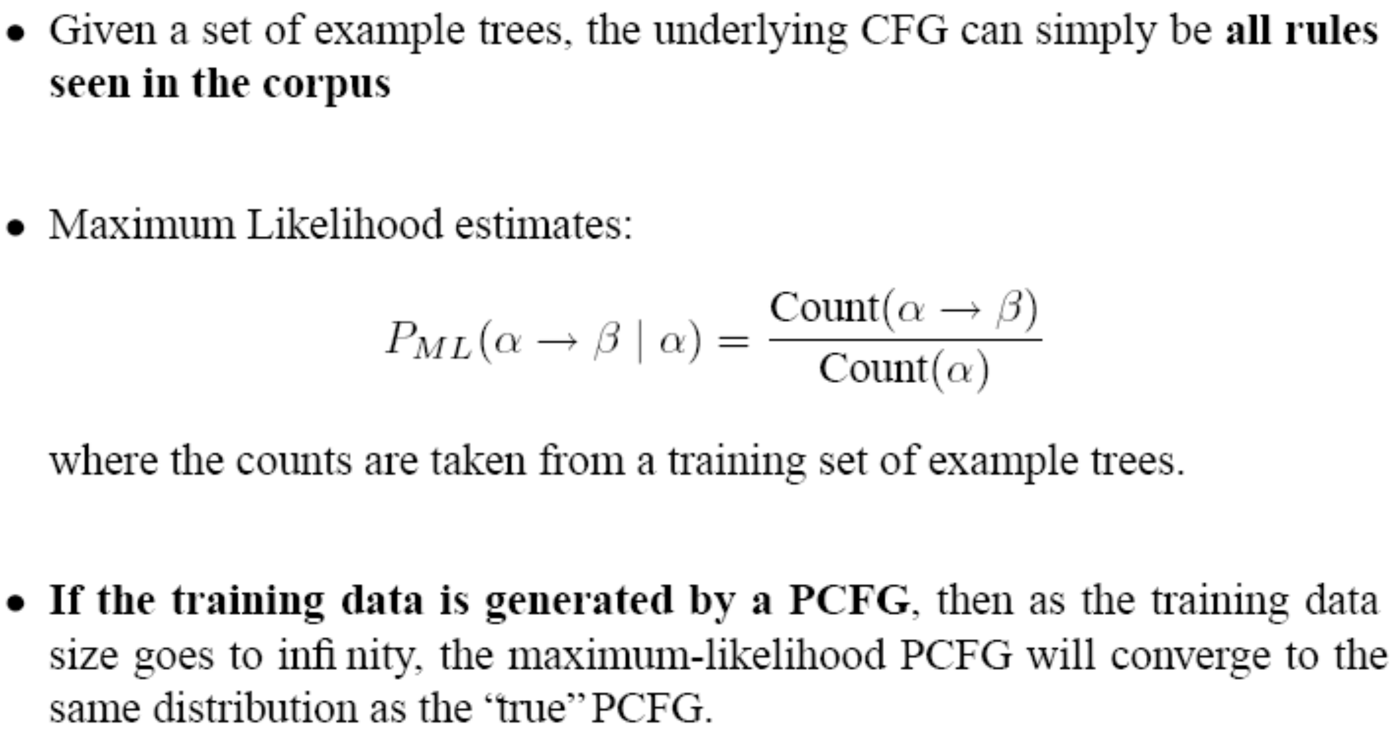
- Assigns a probability to each *left-most derivation*, or parse-tree, allowed by the underlying CFG

- Say we have a sentence S, set of derivations for that sentence is $\mathcal{T}(S)$. Then a PCFG assigns a probability to each member of $\mathcal{T}(S)$. i.e., *we now have a ranking in order of probability.*

- The probability of a string S is

$$\sum_{T \in \mathcal{T}(S)} P(T, S)$$

Deriving a PCFG from a Corpus



- Given a set of example trees, the underlying CFG can simply be **all rules seen in the corpus**

- Maximum Likelihood estimates:

$$P_{ML}(\alpha \rightarrow \beta \mid \alpha) = \frac{\text{Count}(\alpha \rightarrow \beta)}{\text{Count}(\alpha)}$$

  where the counts are taken from a training set of example trees.

- **If the training data is generated by a PCFG**, then as the training data size goes to infinity, the maximum-likelihood PCFG will converge to the same distribution as the ‘true” PCFG.

Algorithms for PCFG

- Given a PCFG and a sentence S, define $\mathcal{T}(S)$ to be the set of trees with S as the yield.

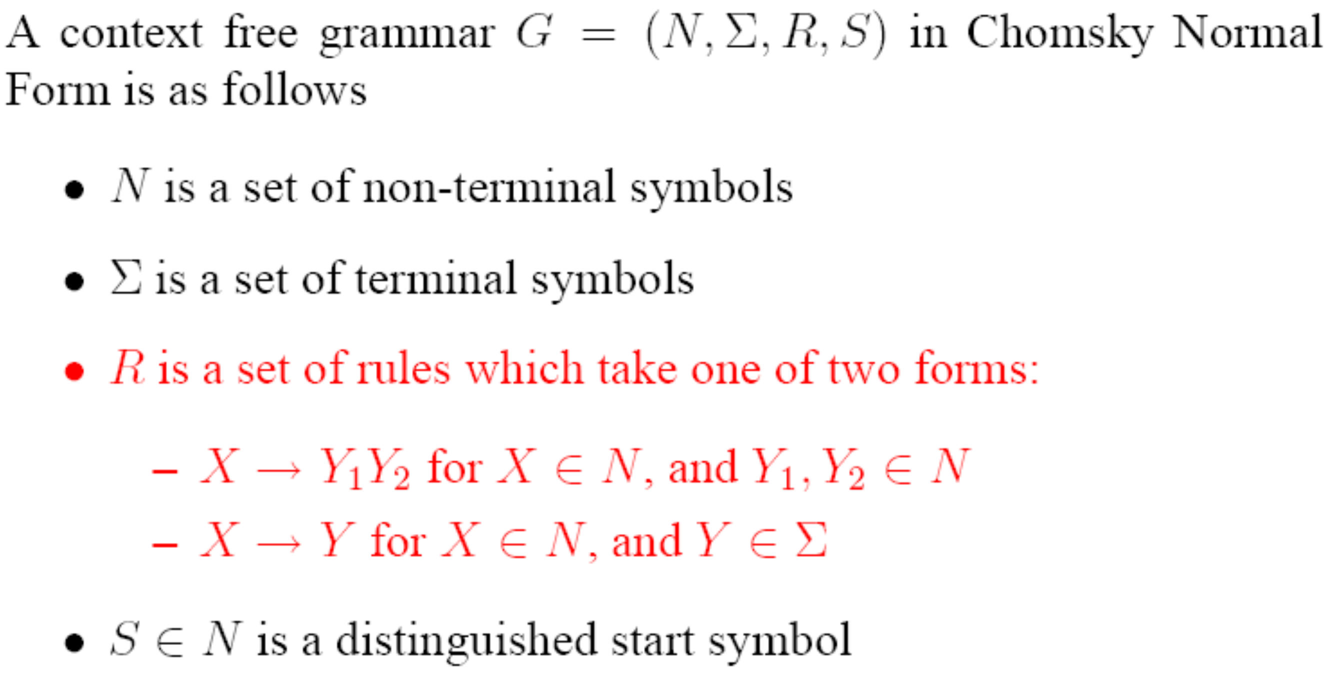
- Given a PCFG and a sentence S, how do we find

$$\arg\max_{T \in \mathcal{T}(S)} P(T, S)$$

- Given a PCFG and a sentence S, how do we find

$$P(S) = \sum_{T \in \mathcal{T}(S)} P(T, S)$$

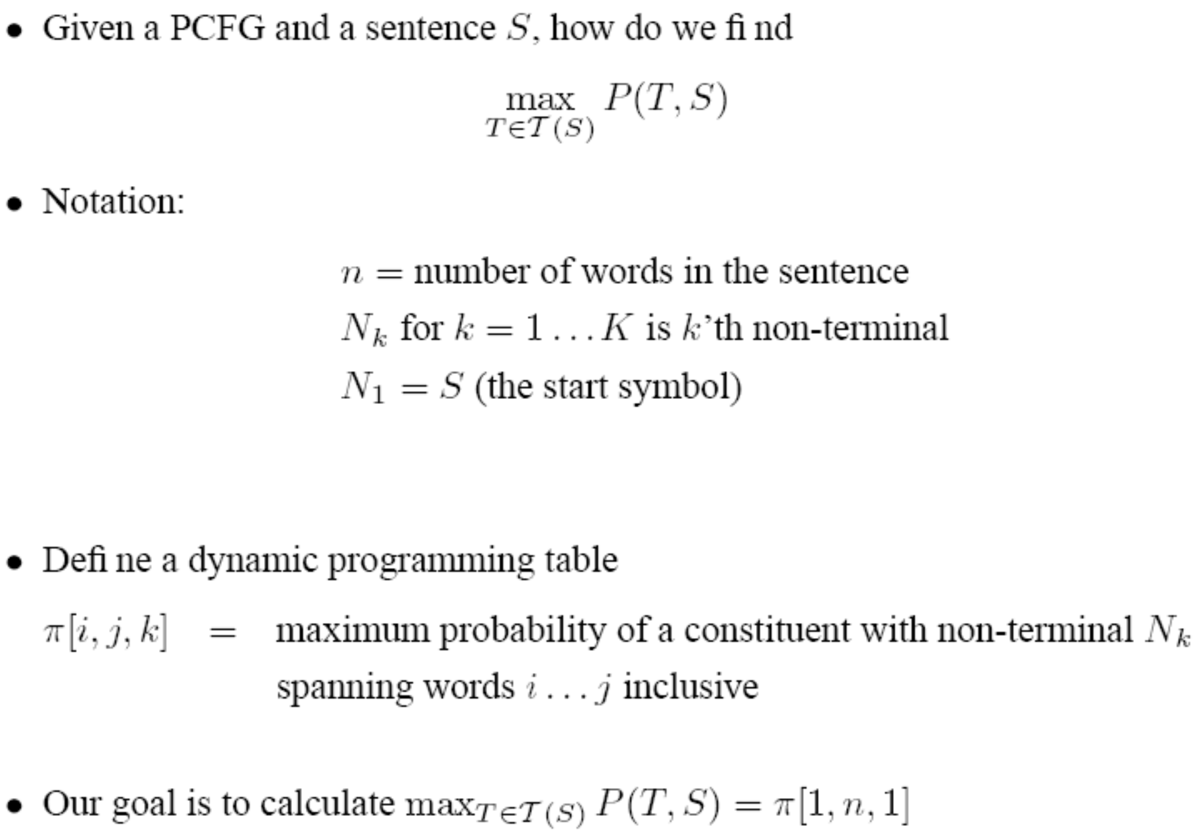
Chomsky Normal Form



A context free grammar $G = (N, \Sigma, R, S)$ in Chomsky Normal Form is as follows

- N is a set of non-terminal symbols
- Σ is a set of terminal symbols
- R is a set of rules which take one of two forms:
  - $X \rightarrow Y_1 Y_2$ for $X \in N$, and $Y_1, Y_2 \in N$
  - $X \rightarrow Y$ for $X \in N$, and $Y \in \Sigma$
- $S \in N$ is a distinguished start symbol

A Dynamic Programming Algorithm for the Max



- Given a PCFG and a sentence S, how do we find

$$\max_{T \in \mathcal{T}(S)} P(T, S)$$

- Notation:

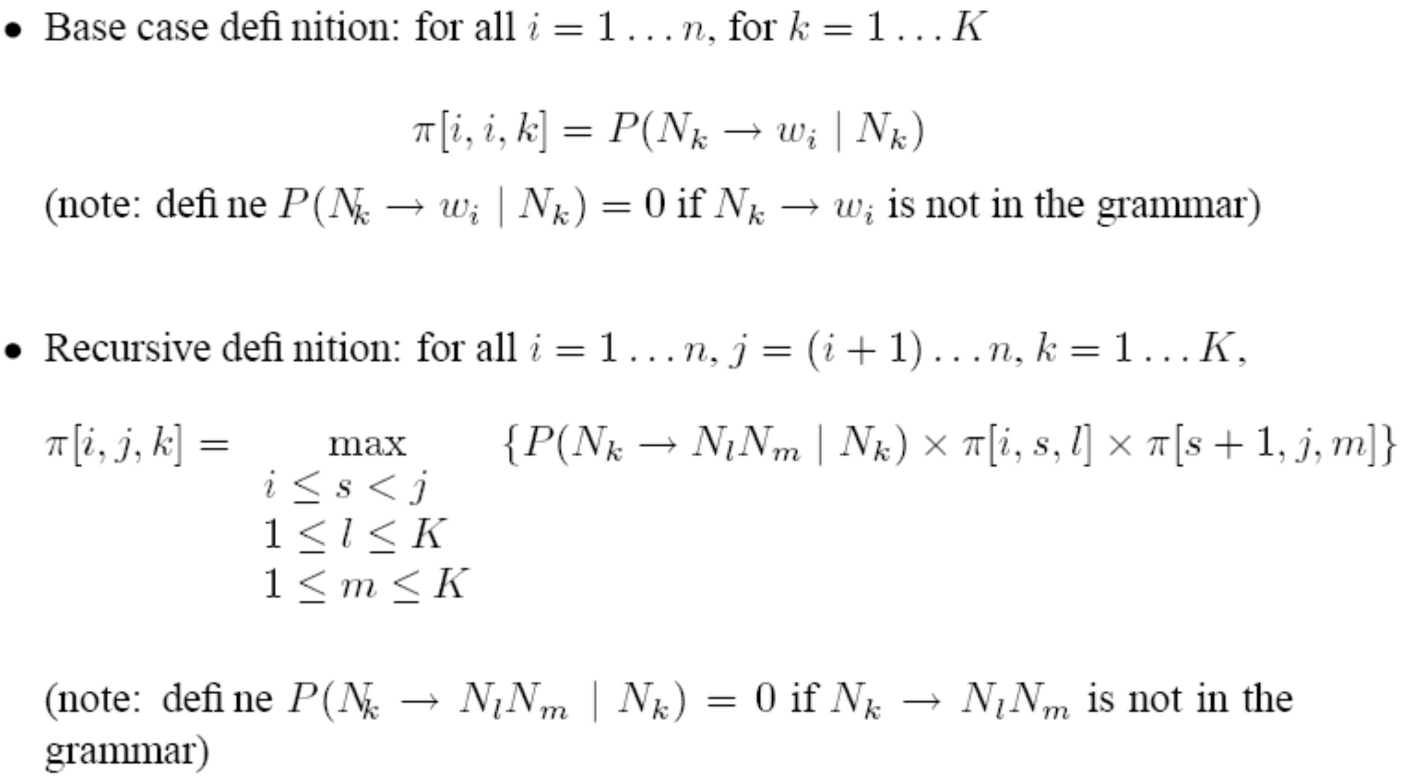
$$n = \text{number of words in the sentence}$$
$$N_k \text{ for } k = 1 \ldots K \text{ is } k\text{'th non-terminal}$$
$$N_1 = S \text{ (the start symbol)}$$

- Define a dynamic programming table

$\pi[i, j, k]$ = maximum probability of a constituent with non-terminal N_k spanning words $i \ldots j$ inclusive

- Our goal is to calculate $\max_{T \in \mathcal{T}(S)} P(T, S) = \pi[1, n, 1]$

A Dynamic Programming Algorithm for the Max (cont.)



- Base case definition: for all $i = 1 \ldots n$, for $k = 1 \ldots K$

$$\pi[i, i, k] = P(N_k \rightarrow w_i \mid N_k)$$

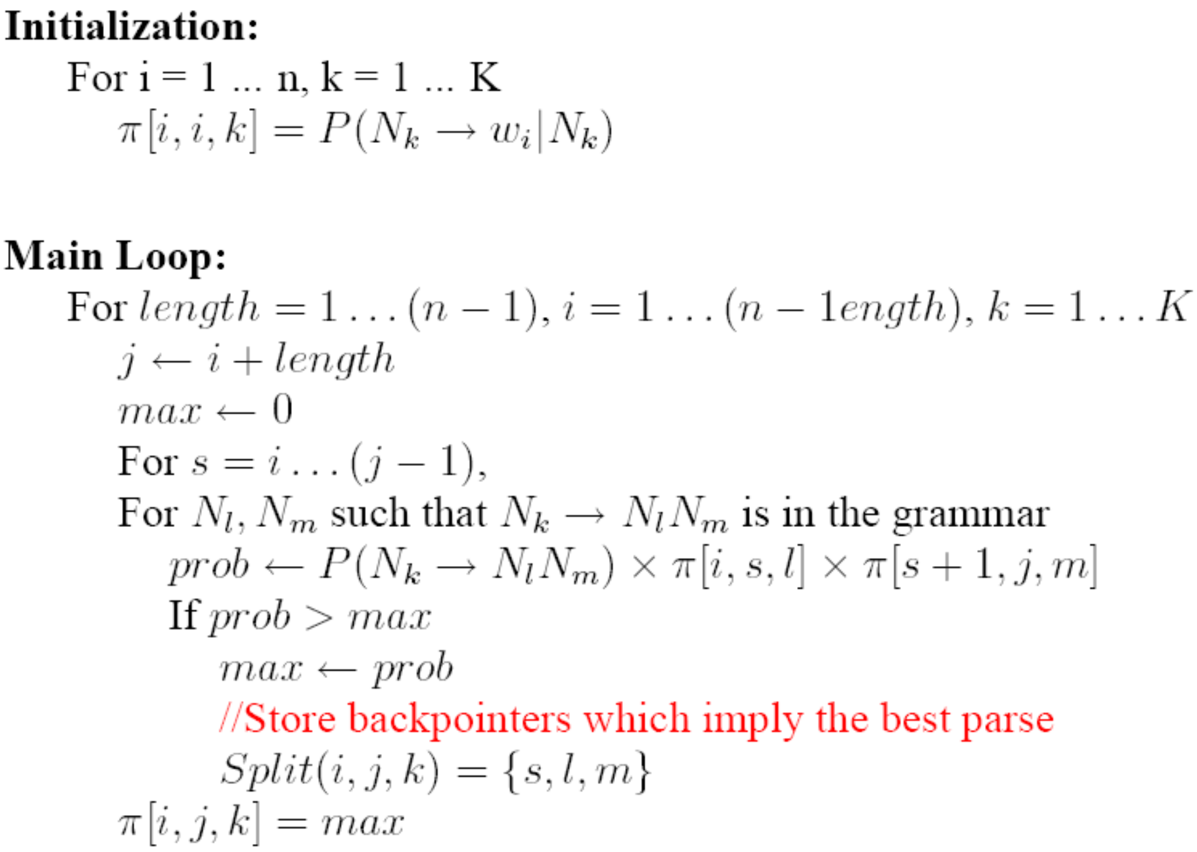
(note: define $P(N_k \rightarrow w_i \mid N_k) = 0$ if $N_k \rightarrow w_i$ is not in the grammar)

- Recursive definition: for all $i = 1 \ldots n$, $j = (i + 1) \ldots n$, $k = 1 \ldots K$,

$$\pi[i, j, k] = \max_{\substack{i \leq s < j \\ 1 \leq l \leq K \\ 1 \leq m \leq K}} \{P(N_k \rightarrow N_l N_m \mid N_k) \times \pi[i, s, l] \times \pi[s + 1, j, m]\}$$

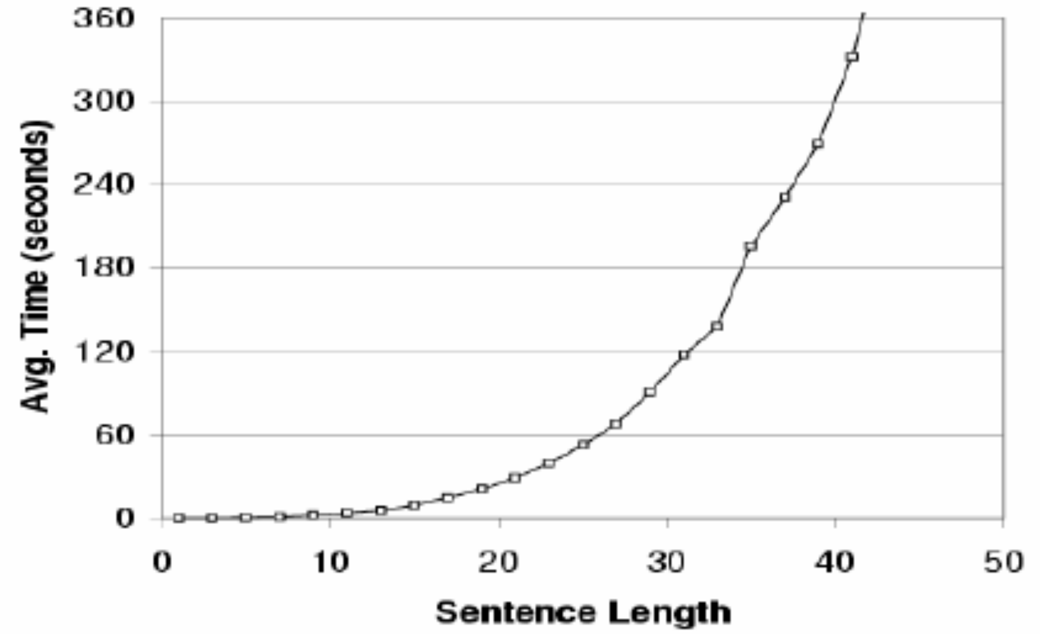
(note: define $P(N_k \rightarrow N_l N_m \mid N_k) = 0$ if $N_k \rightarrow N_l N_m$ is not in the grammar)

A Dynamic Programming Algorithm for the Max (cont.)



```
Initialization:
    For i = 1 ... n, k = 1 ... K
        $\pi[i, i, k] = P(N_k \rightarrow w_i | N_k)$

Main Loop:
    For $length = 1 \ldots (n-1), i = 1 \ldots (n - 1ength), k = 1 \ldots K$
        $j \leftarrow i + length$
        $max \leftarrow 0$
        For $s = i \ldots (j-1)$,
        For $N_l, N_m$ such that $N_k \rightarrow N_l N_m$ is in the grammar
            $prob \leftarrow P(N_k \rightarrow N_l N_m) \times \pi[i, s, l] \times \pi[s+1, j, m]$
            If $prob > max$
                $max \leftarrow prob$
                //Store backpointers which imply the best parse
                $Split(i, j, k) = \{s, l, m\}$
        $\pi[i, j, k] = max$
```

Runtime



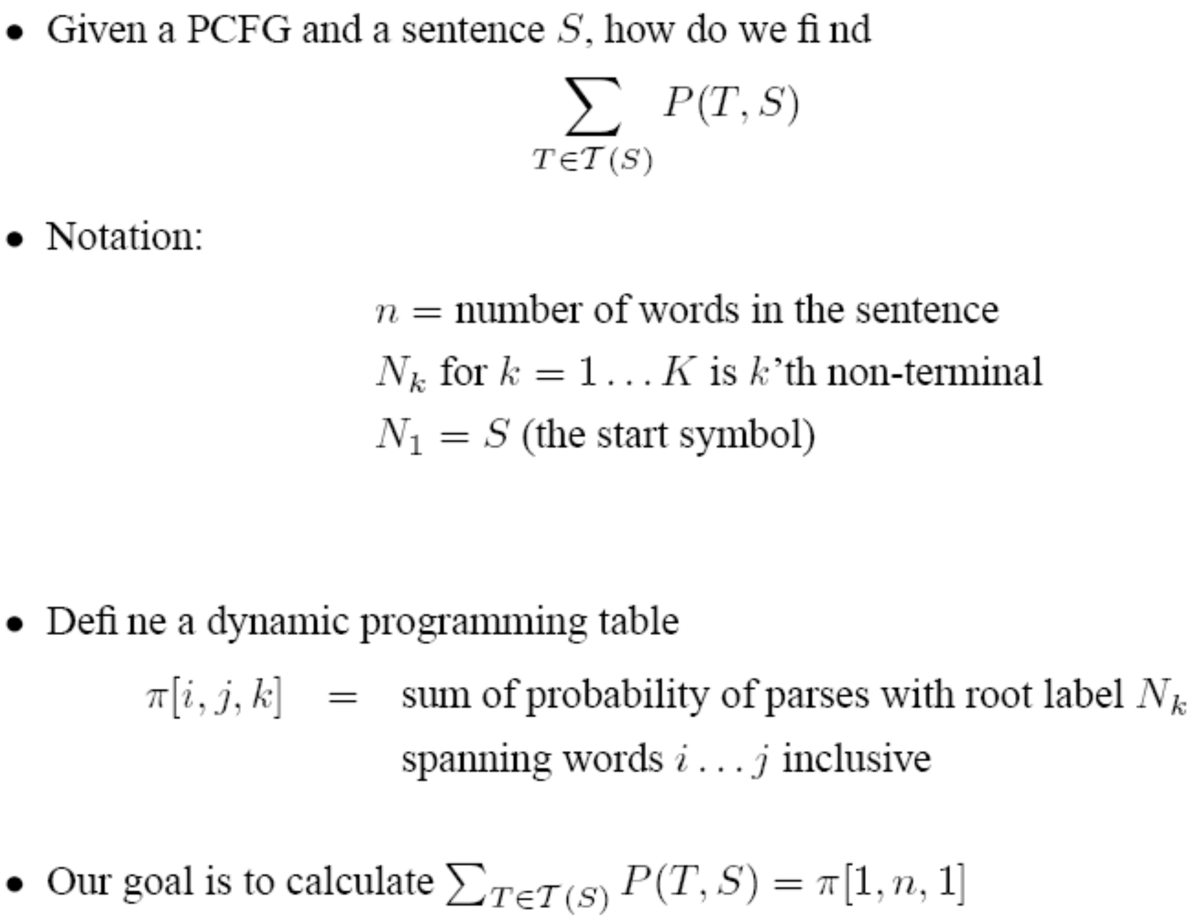
~ 20K Rules

(not an optimized parser!)

Observed exponent:

3.6

A Dynamic Programming Algorithm for the Sum



- Given a PCFG and a sentence S, how do we find

$$\sum_{T \in \mathcal{T}(S)} P(T, S)$$

- Notation:

$n =$ number of words in the sentence

N_k for $k = 1 \ldots K$ is k'th non-terminal

$N_1 = S$ (the start symbol)

- Define a dynamic programming table

$\pi[i, j, k] \quad = \quad$ sum of probability of parses with root label N_k spanning words $i \ldots j$ inclusive

- Our goal is to calculate $\sum_{T \in \mathcal{T}(S)} P(T, S) = \pi[1, n, 1]$

A Dynamic Algorithm for the Sum (cont.)

- Base case definition: for all $i = 1 \ldots n$, for $k = 1 \ldots K$

$$\pi[i, i, k] = P(N_k \rightarrow w_i \mid N_k)$$

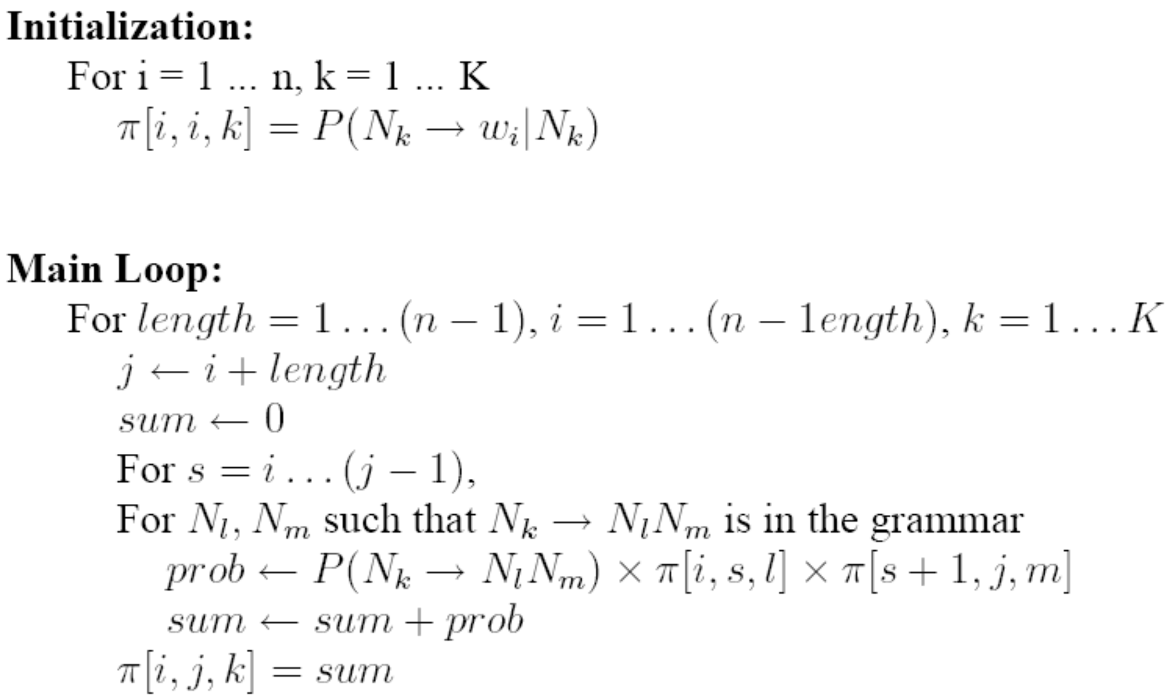
(note: define $P(N_k \rightarrow w_i \mid N_k) = 0$ if $N_k \rightarrow w_i$ is not in the grammar)

- Recursive definition: for all $i = 1 \ldots n$, $j = (i+1) \ldots n$, $k = 1 \ldots K$,

$$\pi[i, j, k] = \sum_{\substack{i \leq s < j \\ 1 \leq l \leq K \\ 1 \leq m \leq K}} \{P(N_k \rightarrow N_l N_m \mid N_k) \times \pi[i, s, l] \times \pi[s+1, j, m]\}$$

(note: define $P(N_k \rightarrow N_l N_m \mid N_k) = 0$ if $N_k \rightarrow N_l N_m$ is not in the grammar)

A Dynamic Programming Algorithm for the Sum (cont.)



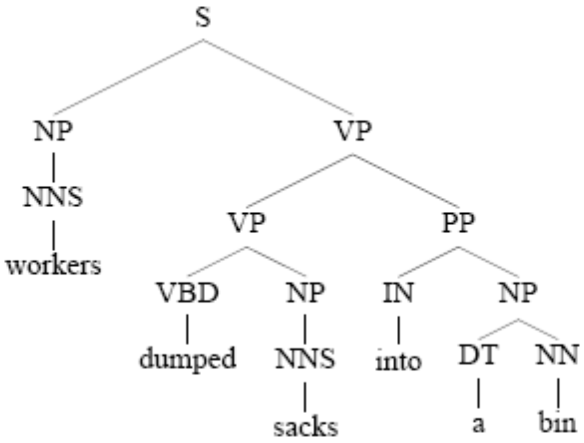
```
**Initialization:**
  For i = 1 ... n, k = 1 ... K
    $\pi[i, i, k] = P(N_k \rightarrow w_i | N_k)$

**Main Loop:**
  For $length = 1 \ldots (n - 1)$, $i = 1 \ldots (n - 1ength)$, $k = 1 \ldots K$
    $j \leftarrow i + length$
    $sum \leftarrow 0$
    For $s = i \ldots (j - 1)$,
    For $N_l, N_m$ such that $N_k \rightarrow N_l N_m$ is in the grammar
      $prob \leftarrow P(N_k \rightarrow N_l N_m) \times \pi[i, s, l] \times \pi[s + 1, j, m]$
      $sum \leftarrow sum + prob$
    $\pi[i, j, k] = sum$
```

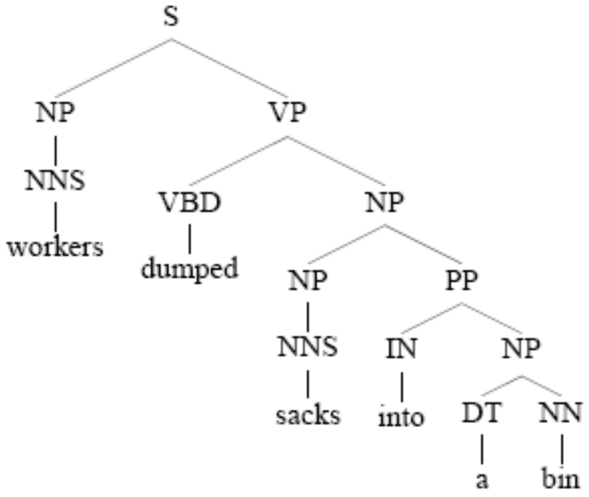
Weaknesses of PCFGs

- Lack of sensitivity to lexical information

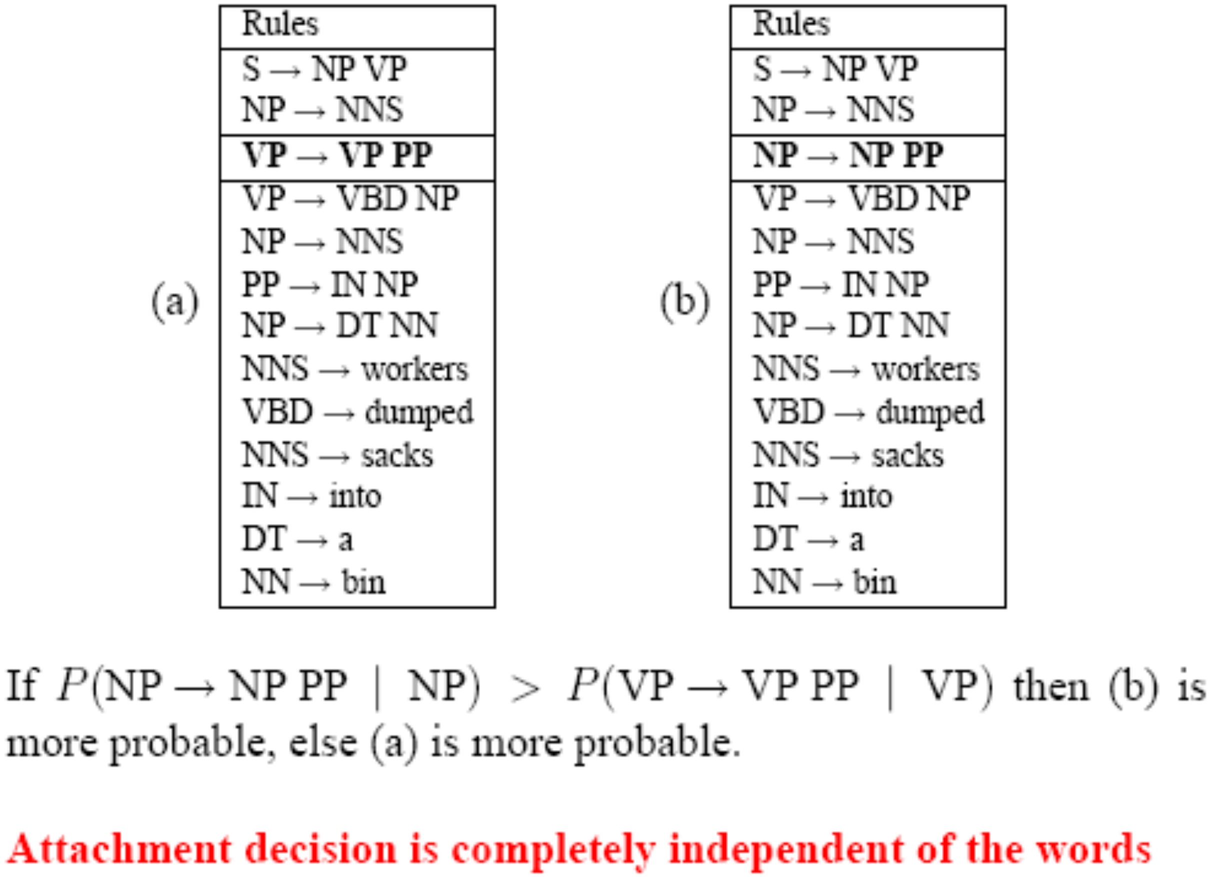
- Lack of sensitivity to structural frequencies

PP Attachment Ambiguity





PP Attachment Ambiguity



(a)

Rules
S → NP VP
NP → NNS
VP → VP PP
VP → VBD NP
NP → NNS
PP → IN NP
NP → DT NN
NNS → workers
VBD → dumped
NNS → sacks
IN → into
DT → a
NN → bin

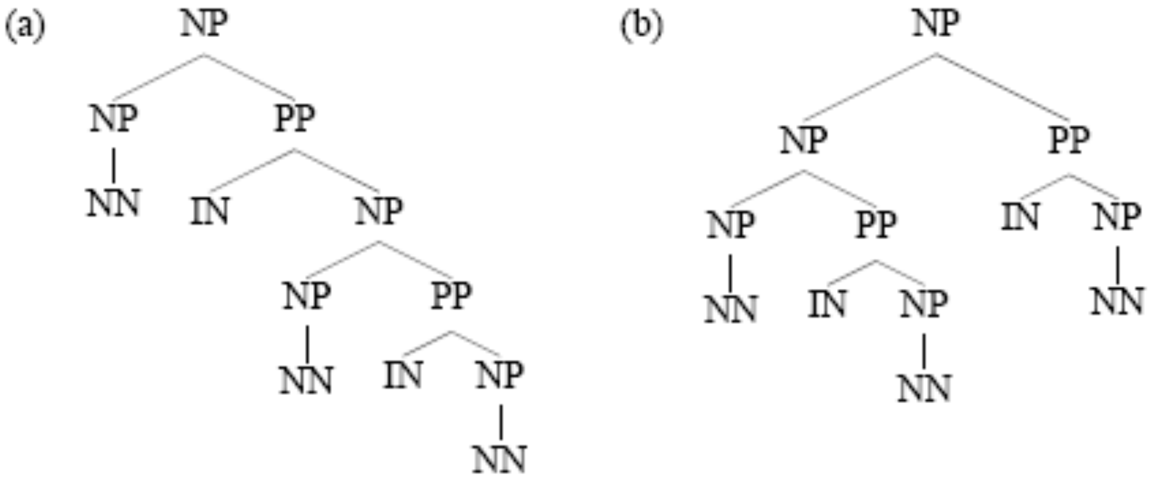
(b)

Rules
S → NP VP
NP → NNS
NP → NP PP
VP → VBD NP
NP → NNS
PP → IN NP
NP → DT NN
NNS → workers
VBD → dumped
NNS → sacks
IN → into
DT → a
NN → bin

If $P(\text{NP} \rightarrow \text{NP PP} \mid \text{NP}) > P(\text{VP} \rightarrow \text{VP PP} \mid \text{VP})$ then (b) is more probable, else (a) is more probable.

Attachment decision is completely independent of the words

Structural Preferences: Close Attachment



- Example: president of a company in Africa

- Both parses have the same rules, therefore receive same probability under a PCFG

- “Close attachment” (structure (a)) is twice as likely in Wall Street Journal text.