Algorithms for NLP

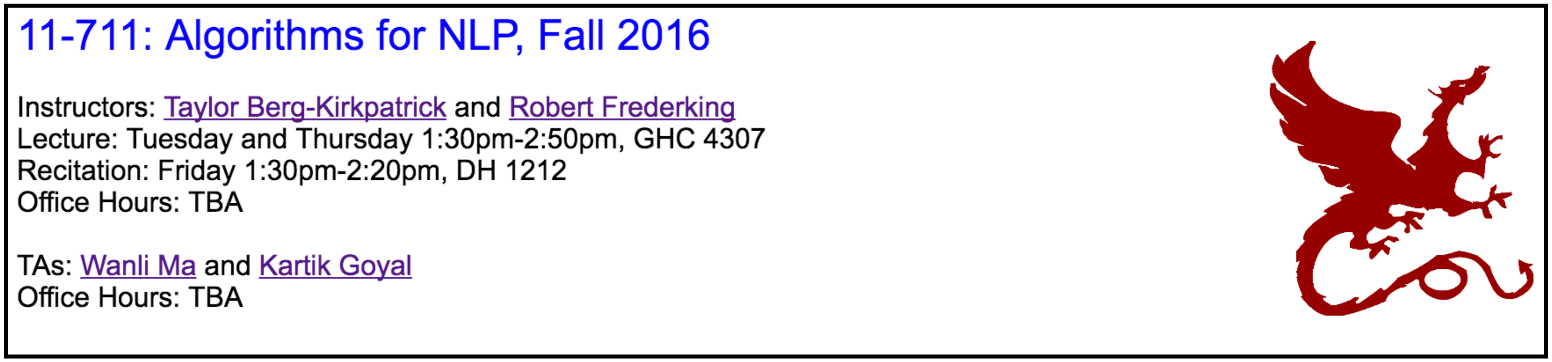
Lecture 1: Introduction

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley

Course Information

http://www.cs.cmu.edu/~tbergkir/11711fa16/

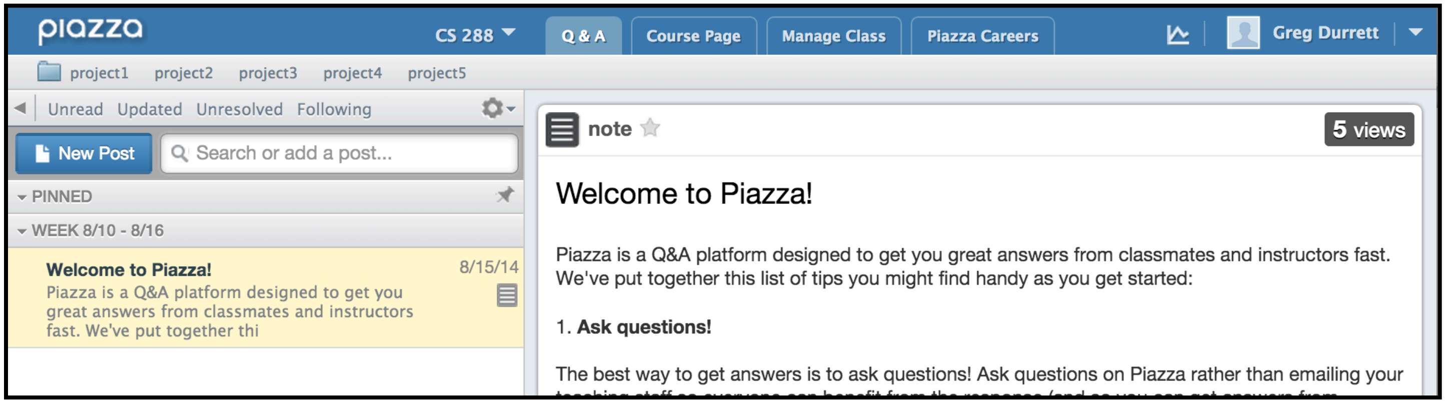


11-711: Algorithms for NLP, Fall 2016

Instructors: Taylor Berg-Kirkpatrick and Robert Frederking
Lecture: Tuesday and Thursday 1:30pm-2:50pm, GHC 4307
Recitation: Friday 1:30pm-2:20pm, DH 1212
Office Hours: TBA

TAs: Wanli Ma and Kartik Goyal
Office Hours: TBA

Piazza setup soon!



Welcome to Piazza! 8/15/14
Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together thi

Welcome to Piazza!

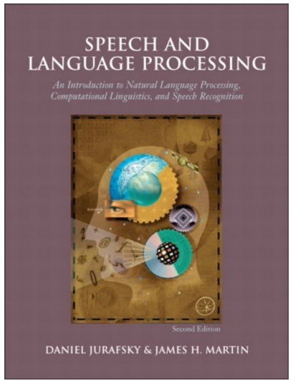
Piazza is a Q&A platform designed to get you great answers from classmates and instructors fast. We've put together this list of tips you might find handy as you get started:

1. **Ask questions!**

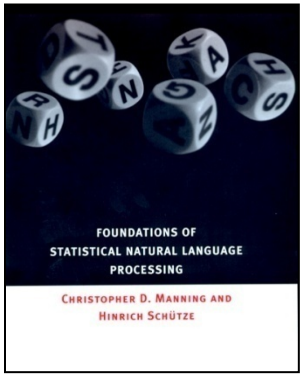
The best way to get answers is to ask questions! Ask questions on Piazza rather than emailing your

Course Requirements

- Prerequisites:
  - Upper division algorithms (dynamic programming)
  - Mastery of basic probability
  - Strong skills in Java or equivalent
  - Deep interest in language


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

- Work and Grading:
  - Four assignments (individual, jars + write-ups)
- Books:
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
  - Also: Manning and Schuetze, Foundations of Statistical NLP


FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

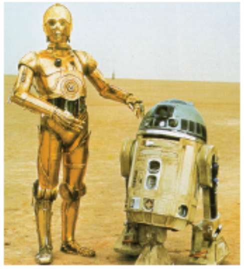
Other Announcements

- Course Contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
- Enrollment: We'll try to take everyone who meets the requirements
- Computing Resources
  - Experiments can take up to hours, even with efficient code
  - Recommendation: start assignments early
- Questions?

AI: Where Do We Stand?

Hollywood

R2D2



KITT

Wall-E

'80 '90 '00 '10

Rule based approaches

Early statistical approaches

Modern statistical approaches

Nimbr 04

Stanford Ra

Reality

Source: Slav Petrov

Language Technologies

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: Deep Understanding

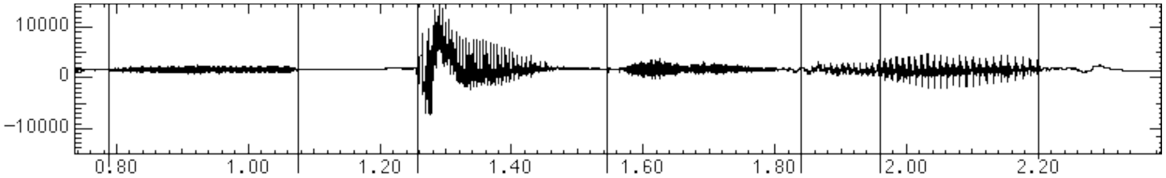
- Requires context, linguistic structure, meanings…

Reality: Shallow Matching

- Requires robustness and scale
- Amazing successes, but fundamental limitations

Speech Systems

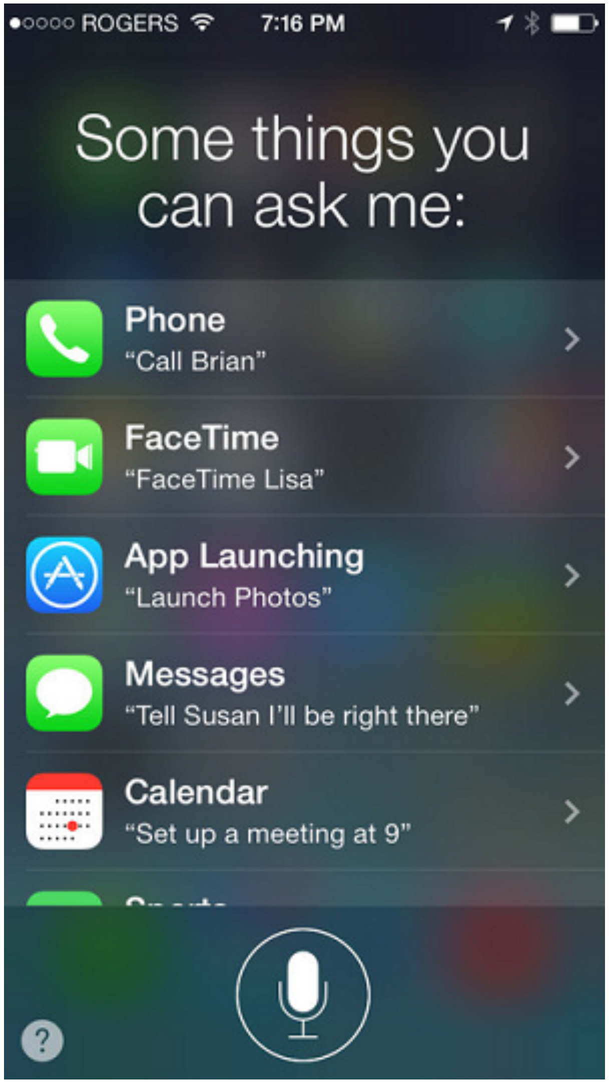
- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV



"Speech Lab"

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)

Example: Siri



●○○○○ ROGERS 7:16 PM

Some things you can ask me:

- Phone
  “Call Brian”
- FaceTime
  “FaceTime Lisa”
- App Launching
  “Launch Photos”
- Messages
  “Tell Susan I’ll be right there”
- Calendar
  “Set up a meeting at 9”
- Sports

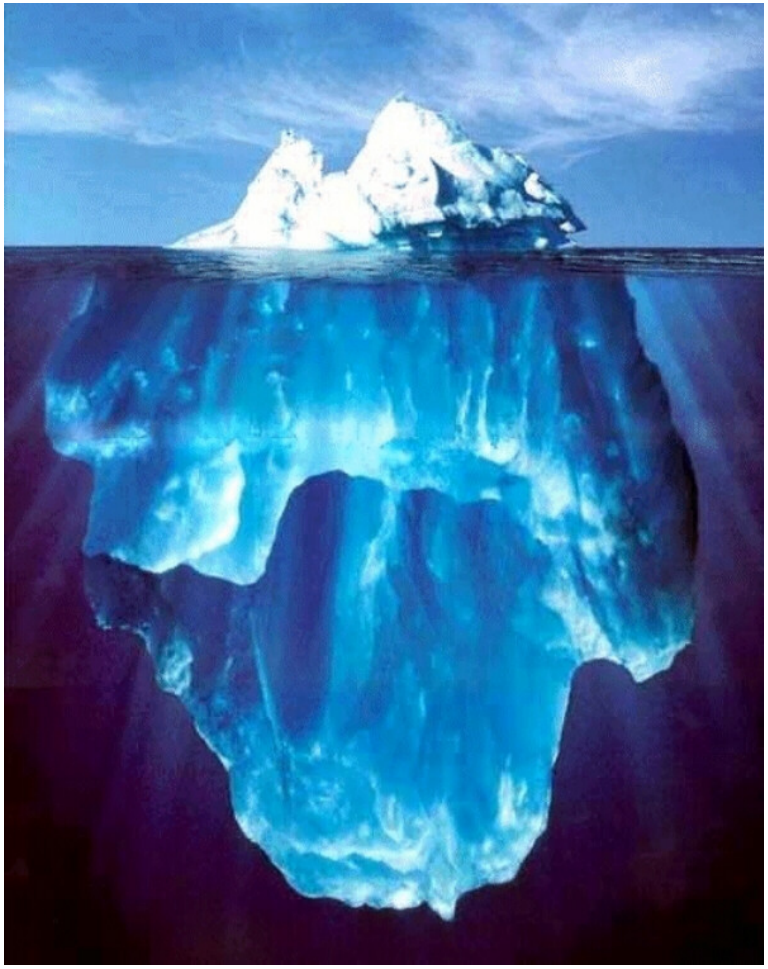
Image: Wikipedia

- Siri contains
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

Text Data is Superficial

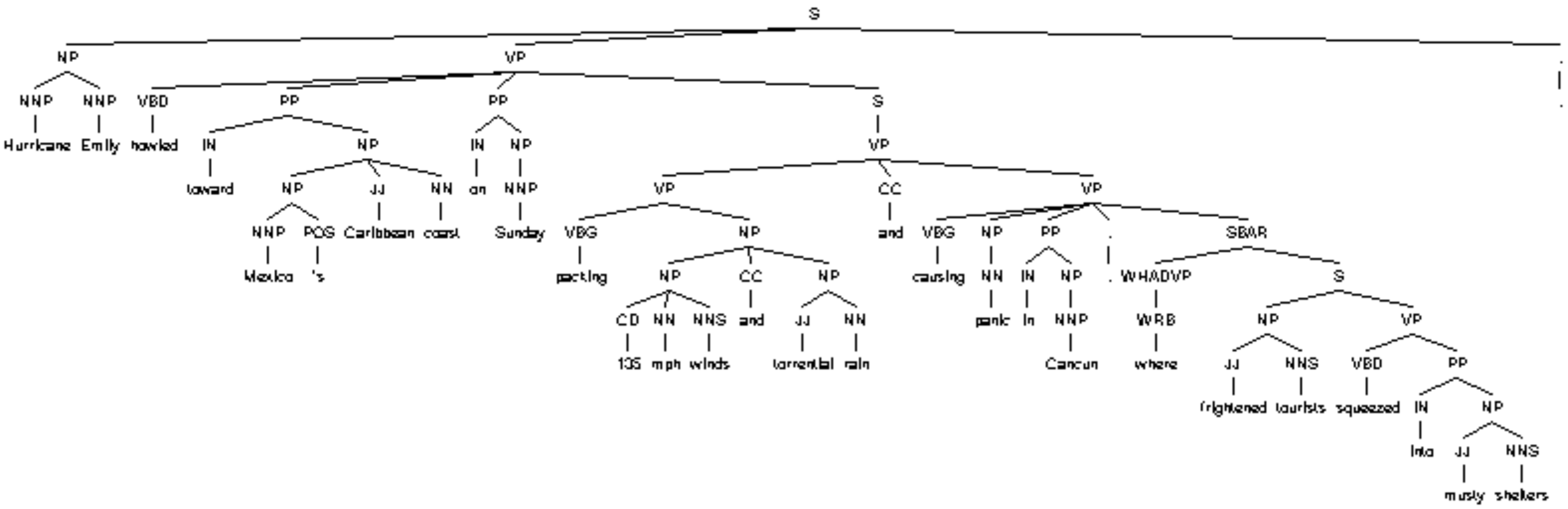
An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

… But Language is Complex



An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

Syntactic Analysis

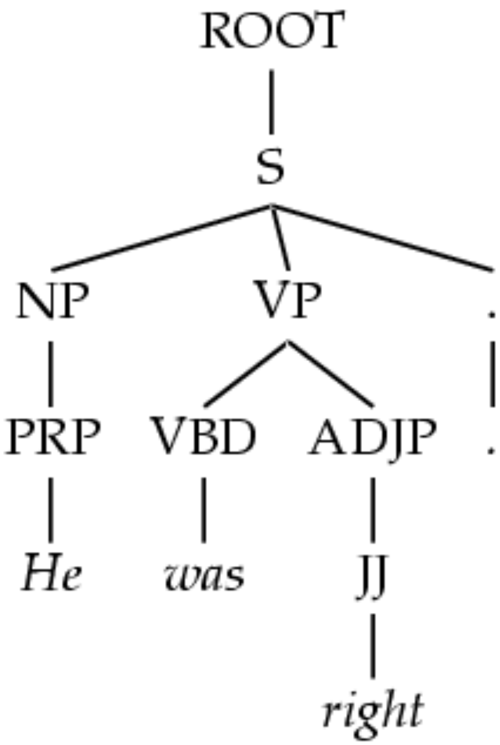


Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- SOTA: ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

Corpus-Based Methods

- A corpus like a treebank gives us three important tools:
  - It gives us broad coverage

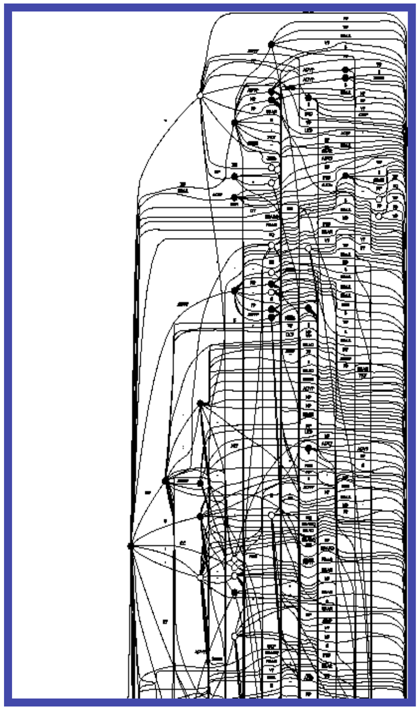


ROOT → S

S → NP VP .

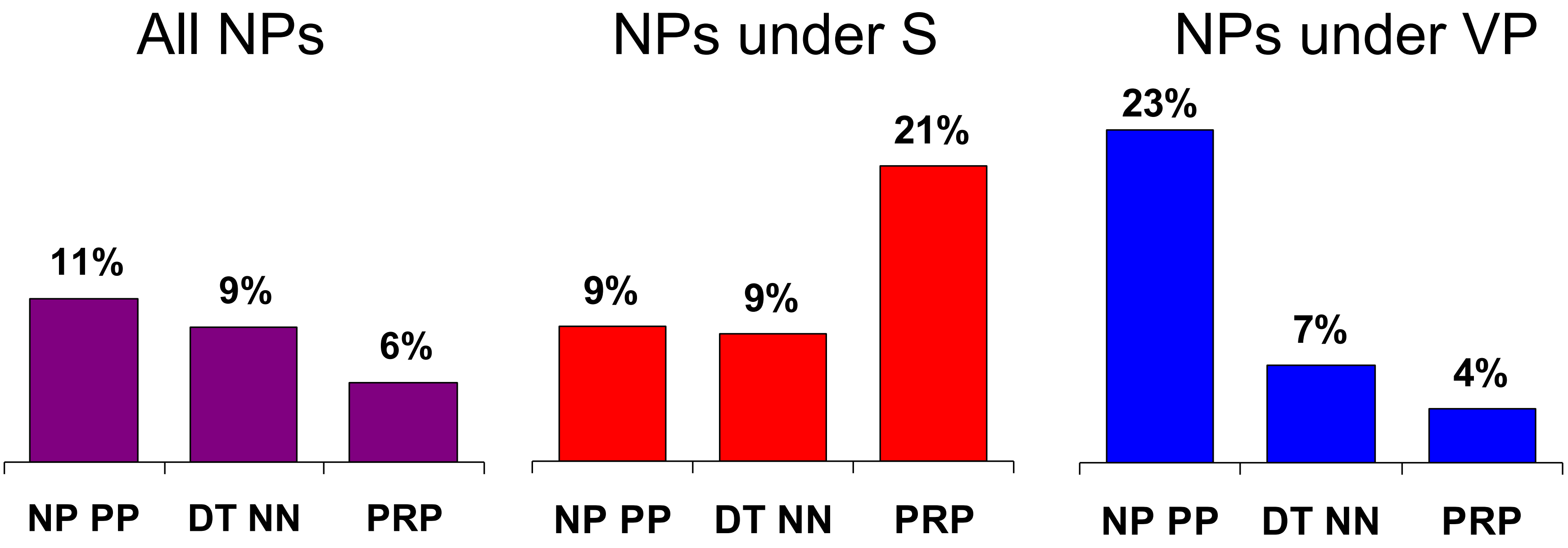
NP → PRP

VP → VBD ADJ



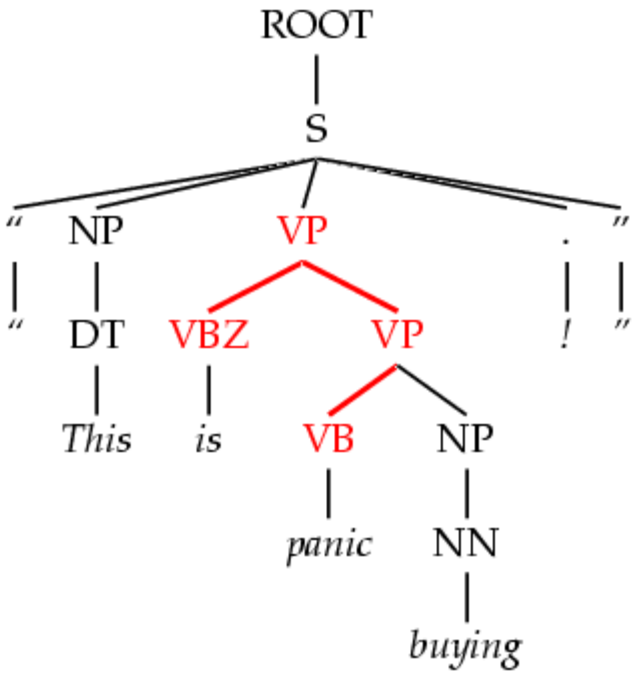
Corpus-Based Methods

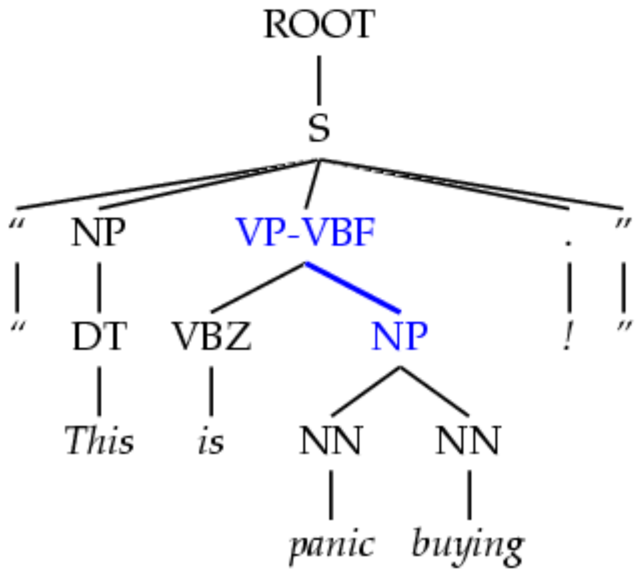
- It gives us statistical information



Corpus-Based Methods

- It lets us check our answers





Semantic Ambiguity

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don’t fully nail down the meaning

I haven’t slept for ten days

John’s boss said he was doing better

- In general, every level of linguistic structure comes with its own ambiguities…

Other Levels of Language

- Tokenization/morphology:
  - What are the words, what is the sub-word structure?
  - Often simple rules work (period after "Mr." isn't sentence break)
  - Relatively easy in English, other languages are harder:
    - Segementation

      哲学家维特根斯坦出生于维也纳

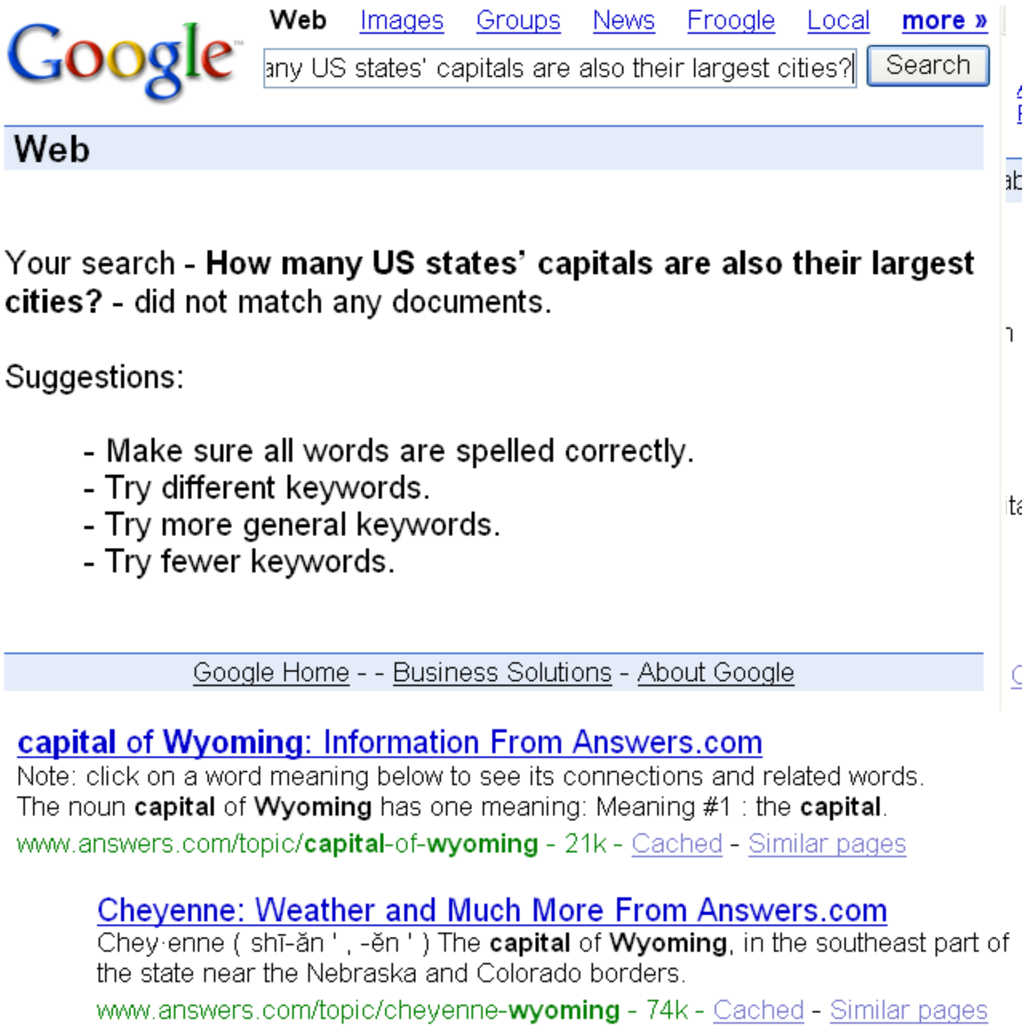
    - Morphology

sarà	*andata*
be+fut+3sg	go+ppt+fem
"she will have gone"	

- Discourse: how do sentences relate to each other?
- Pragmatics: what intent is expressed by the literal meaning, how to react to an utterance?
- Phonetics: acoustics and physical production of sounds
- Phonology: how sounds pattern in a language

Question Answering

- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: “What’s the capital of Wyoming?”
  - Can be harder: “How many US states’ capitals are also their largest cities?”
  - Can be open ended: “What are the main issues in the global warming debate?”
- SOTA: Can do factoids, even when text isn’t a perfect match



any US states' capitals are also their largest cities?

Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.

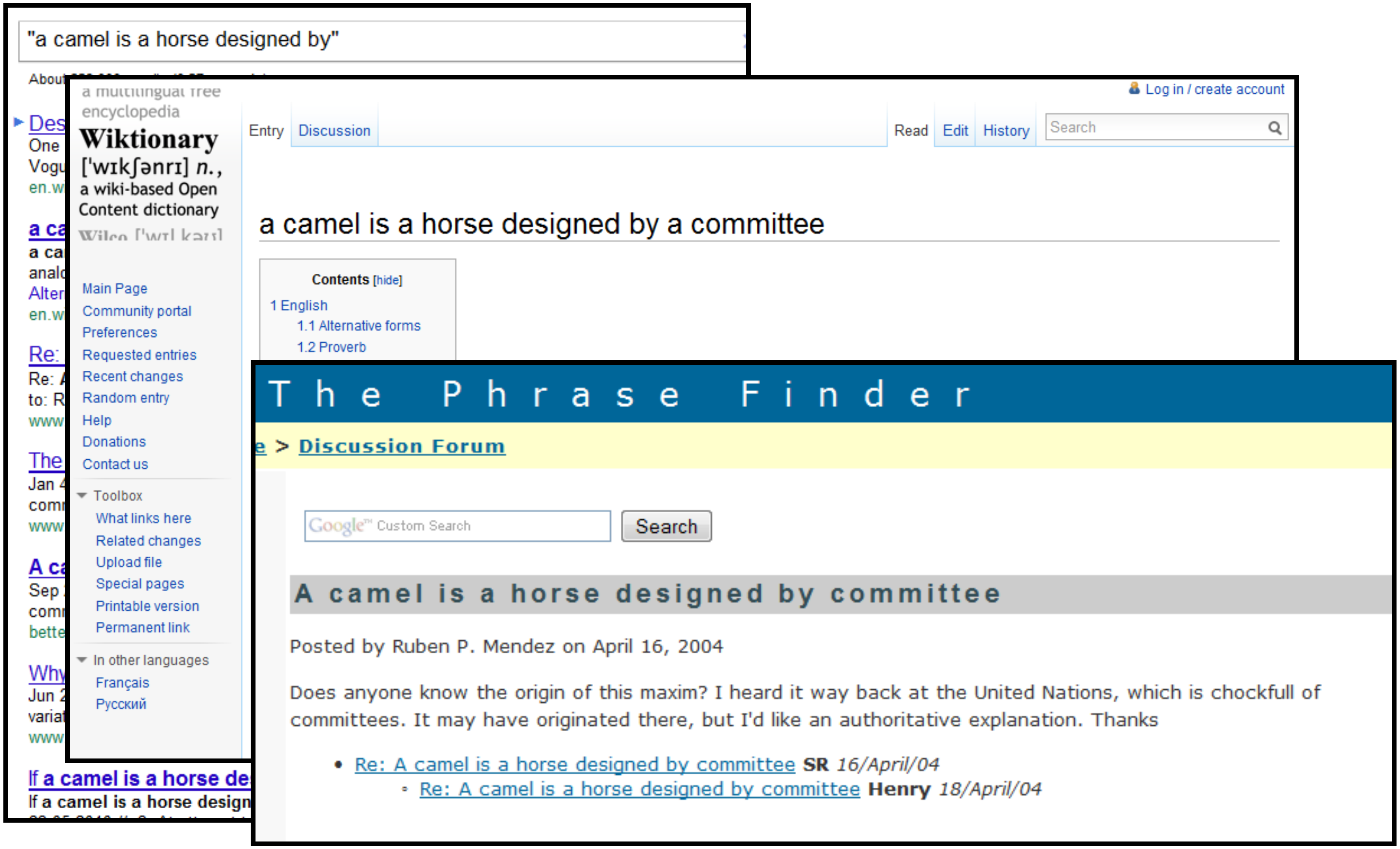
Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

capital of **Wyoming**: Information From Answers.com
Note: click on a word meaning below to see its connections and related words. The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.
www.answers.com/topic/**capital**-of-**wyoming** - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
Chey·enne (shī-ăn ' , -ĕn ') The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-**wyoming** - 74k - Cached - Similar pages

Example: Watson



"a camel is a horse designed by"

About

Des
One
Vogu
en.w

a ca
a ca
analo
Alter
en.w

Re:
Re: A
to: R
www

The
Jan 4
comm
www

A ca
Sep
comm
bette

Why
Jun 2
variat
www

If a camel is a horse de
If a camel is a horse design

a multilingual free encyclopedia

Wiktionary
['wɪkʃənrɪ] n.,
a wiki-based Open Content dictionary

Main Page
Community portal
Preferences
Requested entries
Recent changes
Random entry
Help
Donations
Contact us

Toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link

In other languages
Français
Русский

Log in / create account

Entry | Discussion | Read | Edit | History | Search

a camel is a horse designed by a committee

Contents [hide]

1 English
 1.1 Alternative forms
 1.2 Proverb

The Phrase Finder

e > Discussion Forum

Google™ Custom Search | Search

A camel is a horse designed by committee

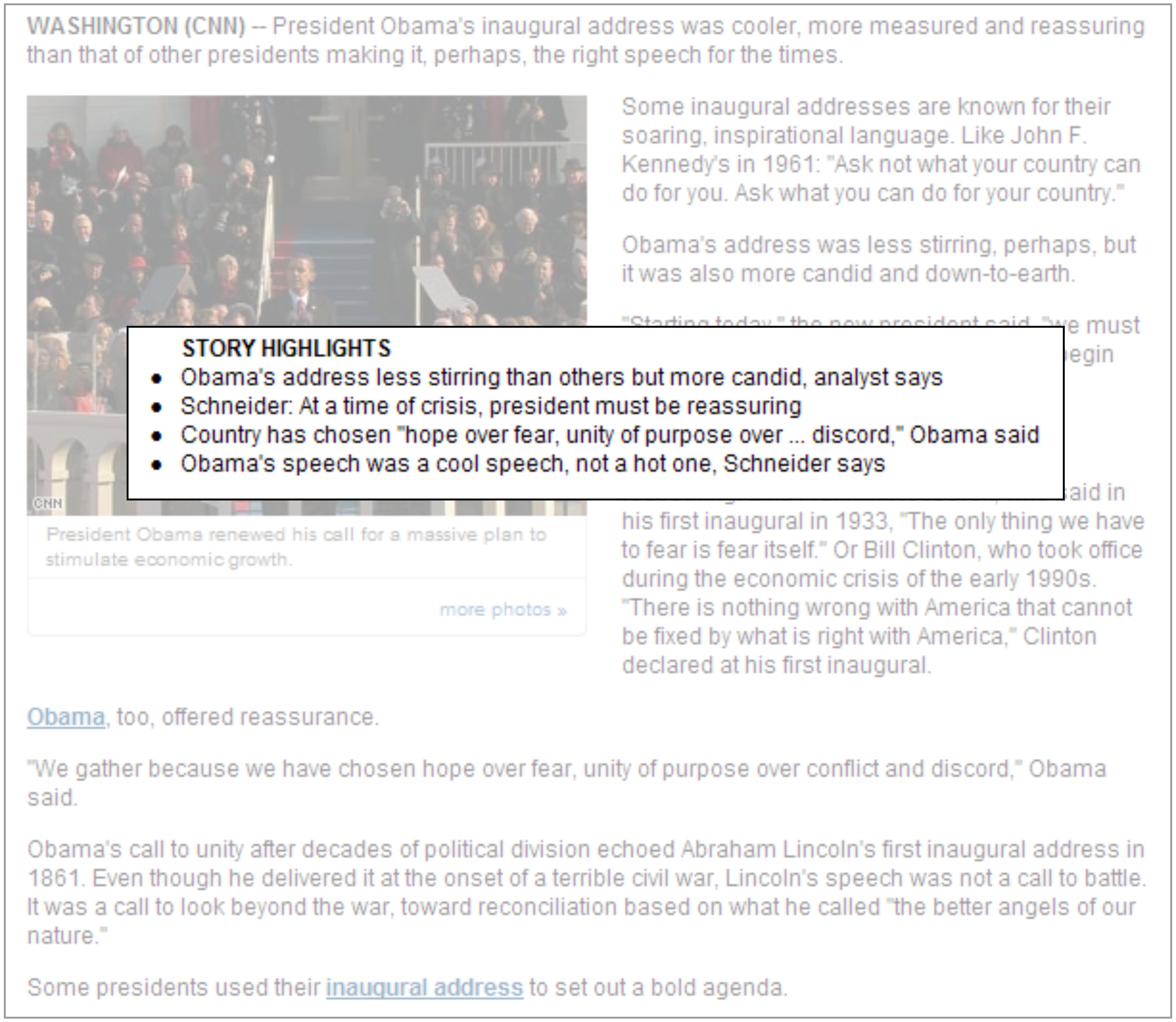
Posted by Ruben P. Mendez on April 16, 2004

Does anyone know the origin of this maxim? I heard it way back at the United Nations, which is chockfull of committees. It may have originated there, but I'd like an authoritative explanation. Thanks

- Re: A camel is a horse designed by committee **SR** *16/April/04*
  - Re: A camel is a horse designed by committee **Henry** *18/April/04*

Summarization

- Condensing documents
- An example of analysis with generation



WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must

egin

STORY HIGHLIGHTS

- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

aid in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

Obama, too, offered reassurance.

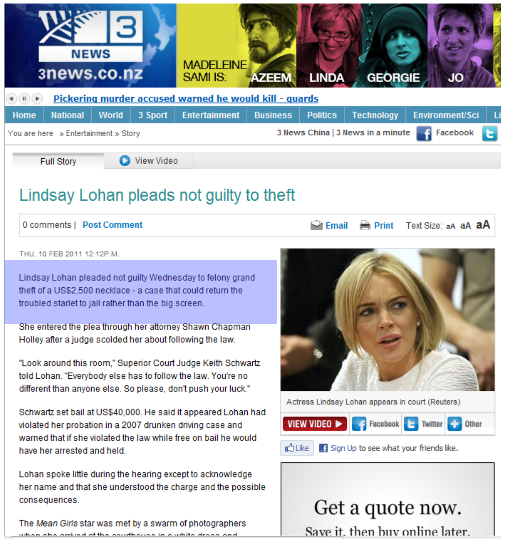
"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their inaugural address to set out a bold agenda.

Extractive Summaries

Lindsay Lohan pleaded not guilty Wednesday to felony grand theft of a $2,500 necklace, a case that could return the troubled starlet to jail rather than the big screen. Saying it appeared that Lohan had violated her probation in a 2007 drunken driving case, the judge set bail at $40,000 and warned that if Lohan was accused of breaking the law while free he would have her held without bail. The Mean Girls star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early.


NEWS
3news.co.nz
MADELEINE SAMI IS: AZEEM LINDA GEORGIE JO
Pickering murder accused warned he would kill - guards
Home National World 3 Sport Entertainment Business Politics Technology Environment/Sci
You are here » Entertainment » Story
3 News China | 3 News in a minute Facebook
Full Story View Video
0 comments | Post Comment Email Print Text Size: aA aA aA
VIEW VIDEO Facebook Twitter Other
Like Sign Up to see what your friends like.
Get a quote now.
Save it, then buy online later.

Lindsay Lohan pleads not guilty to theft

THU, 10 FEB 2011 12:12P.M.

Lindsay Lohan pleaded not guilty Wednesday to felony grand theft of a US$2,500 necklace - a case that could return the troubled starlet to jail rather than the big screen.

She entered the plea through her attorney Shawn Chapman Holley after a judge scolded her about following the law.

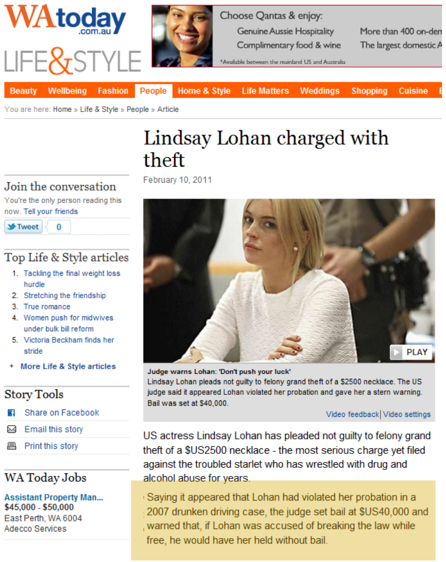
"Look around this room," Superior Court Judge Keith Schwartz told Lohan. "Everybody else has to follow the law. You're no different than anyone else. So please, don't push your luck."

Schwartz set bail at US$40,000. He said it appeared Lohan had violated her probation in a 2007 drunken driving case and warned that if she violated the law while free on bail he would have her arrested and held.

Lohan spoke little during the hearing except to acknowledge her name and that she understood the charge and the possible consequences.

The *Mean Girls* star was met by a swarm of photographers

Actress Lindsay Lohan appears in court (Reuters)


WAtoday.com.au
LIFE&STYLE
Choose Qantas & enjoy:
Genuine Aussie Hospitality
Complimentary food & wine
More than 400 on-den
The largest domestic A
*Available between the mainland US and Australia
Beauty Wellbeing Fashion People Home & Style Life Matters Weddings Shopping Cuisine
You are here: Home » Life & Style » People » Article
PLAY
Tweet 0

Lindsay Lohan charged with theft

February 10, 2011

Judge warns Lohan: 'Don't push your luck'
Lindsay Lohan pleads not guilty to felony grand theft of a $2500 necklace. The US judge said it appeared Lohan violated her probation and gave her a stern warning. Bail was set at $40,000.

Video feedback | Video settings

US actress Lindsay Lohan has pleaded not guilty to felony grand theft of a $US2500 necklace - the most serious charge yet filed against the troubled starlet who has wrestled with drug and alcohol abuse for years.

Saying it appeared that Lohan had violated her probation in a 2007 drunken driving case, the judge set bail at $US40,000 and warned that, if Lohan was accused of breaking the law while free, he would have her held without bail.

Join the conversation

You're the only person reading this now. Tell your friends

Top Life & Style articles

1. Tackling the final weight loss hurdle
2. Stretching the friendship
3. True romance
4. Women push for midwives under bulk bill reform
5. Victoria Beckham finds her stride

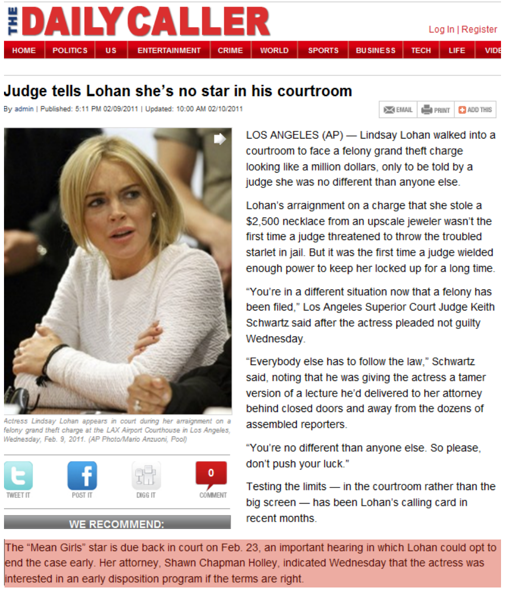
+ More Life & Style articles

Story Tools

Share on Facebook
Email this story
Print this story

WA Today Jobs

Assistant Property Man...
$45,000 - $50,000
East Perth, WA 6004
Adecco Services


THE DAILY CALLER
Log In | Register
HOME POLITICS US ENTERTAINMENT CRIME WORLD SPORTS BUSINESS TECH LIFE VIDE
EMAIL PRINT ADD THIS
TWEET IT POST IT DIGG IT 0 COMMENT
WE RECOMMEND:

Judge tells Lohan she's no star in his courtroom

By admin | Published: 5:11 PM 02/09/2011 | Updated: 10:00 AM 02/10/2011

LOS ANGELES (AP) — Lindsay Lohan walked into a courtroom to face a felony grand theft charge looking like a million dollars, only to be told by a judge she was no different than anyone else.

Lohan's arraignment on a charge that she stole a $2,500 necklace from an upscale jeweler wasn't the first time a judge threatened to throw the troubled starlet in jail. But it was the first time a judge wielded enough power to keep her locked up for a long time.

"You're in a different situation now that a felony has been filed," Los Angeles Superior Court Judge Keith Schwartz said after the actress pleaded not guilty Wednesday.

"Everybody else has to follow the law," Schwartz said, noting that he was giving the actress a tamer version of a lecture he'd delivered to her attorney behind closed doors and away from the dozens of assembled reporters.

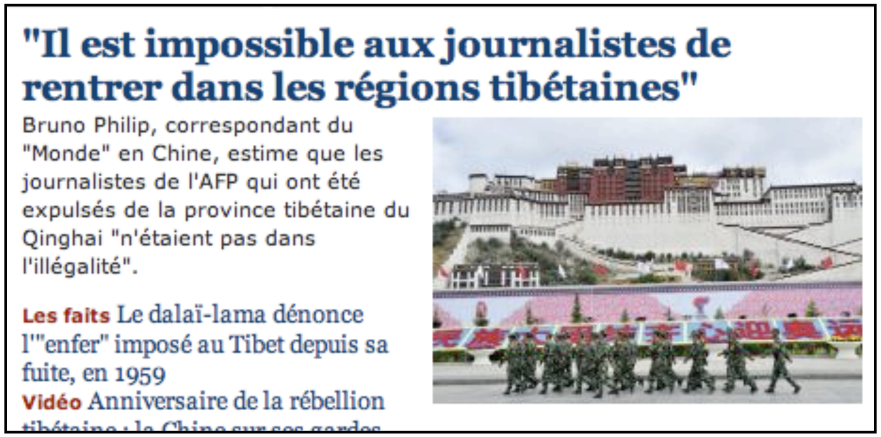
"You're no different than anyone else. So please, don't push your luck."

Testing the limits — in the courtroom rather than the big screen — has been Lohan's calling card in recent months.

The "Mean Girls" star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early. Her attorney, Shawn Chapman Holley, indicated Wednesday that the actress was interested in an early disposition program if the terms are right.

Actress Lindsay Lohan appears in court during her arraignment on a felony grand theft charge at the LAX Airport Courthouse in Los Angeles, Wednesday, Feb. 9, 2011. (AP Photo/Mario Anzuoni, Pool)

Machine Translation

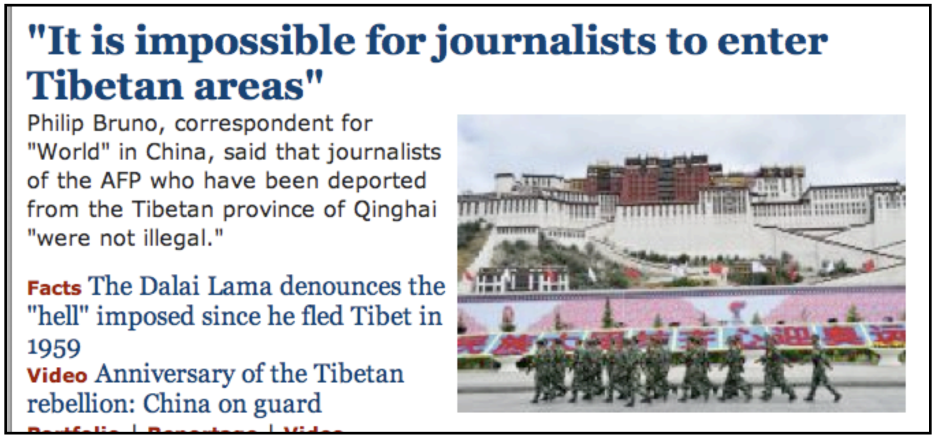


"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959

Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes



"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

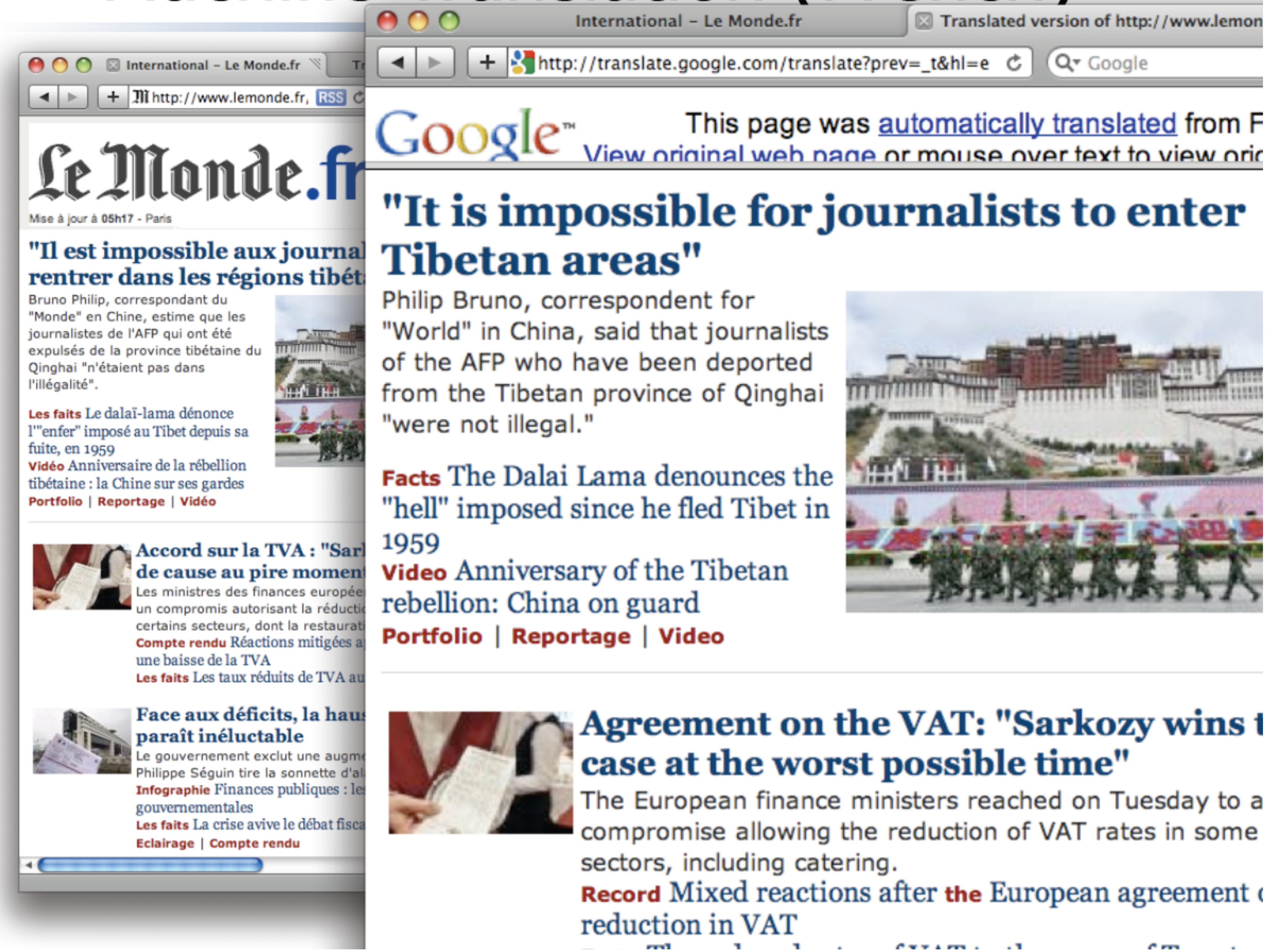
Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

Video Anniversary of the Tibetan rebellion: China on guard

Portfolio | Reportage | Video

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)

Machine Translation (French)



Le Monde.fr

Mise à jour à 05h17 - Paris

"Il est impossible aux journal
rentrer dans les régions tibét

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959

Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes

Portfolio | Reportage | Vidéo

Accord sur la TVA : "Sar
de cause au pire momen

Les ministres des finances europée
un compromis autorisant la réducti
certains secteurs, dont la restaurat

Compte rendu Réactions mitigées a
une baisse de la TVA

Les faits Les taux réduits de TVA au

Face aux déficits, la haus
paraît inéluctable

Le gouvernement exclut une augm
Philippe Séguin tire la sonnette d'al

Infographie Finances publiques : le
gouvernementales

Les faits La crise avive le débat fisca

Eclairage | Compte rendu

International – Le Monde.fr
Translated version of http://www.lemon
http://translate.google.com/translate?prev=_t&hl=e
Google

Google™ This page was automatically translated from F
View original web page or mouse over text to view orig

"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

Video Anniversary of the Tibetan rebellion: China on guard

Portfolio | Reportage | Video

Agreement on the VAT: "Sarkozy wins t
case at the worst possible time"

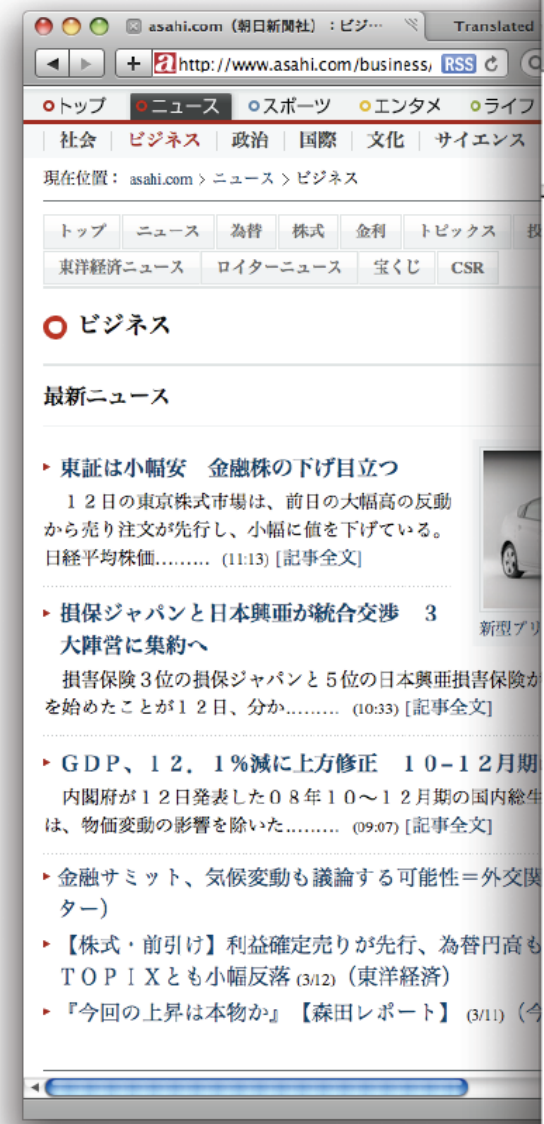
The European finance ministers reached on Tuesday to a compromise allowing the reduction of VAT rates in some sectors, including catering.

Record Mixed reactions after the European agreement o
reduction in VAT

More Data: Machine Translation

SOURCE	Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.
HUMAN	That would be an interim solution which would make it possible to work towards a binding charter in the long term .
1x DATA	[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]
10x DATA	[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]
100x DATA	[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]
1000x DATA	[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]

Machine Translation (Japanese)



ビジネス

最新ニュース

- 東証は小幅安　金融株の下げ目立つ

　１２日の東京株式市場は、前日の大幅高の反動から売り注文が先行し、小幅に値を下げている。日経平均株価……… (11:13) [記事全文]

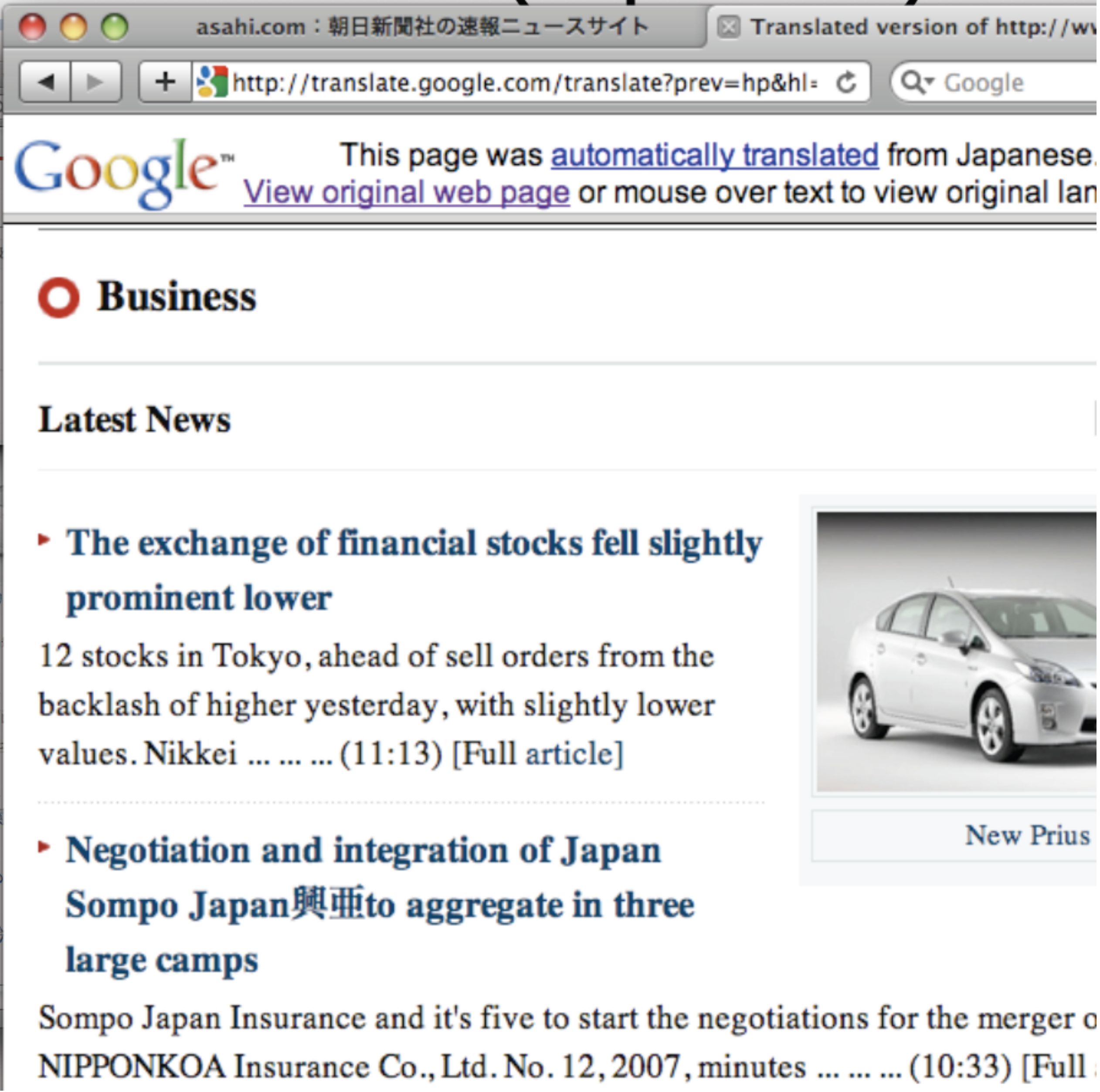
- 損保ジャパンと日本興亜が統合交渉　３大陣営に集約へ

　損害保険３位の損保ジャパンと５位の日本興亜損害保険が を始めたことが１２日、分か……… (10:33) [記事全文]

- ＧＤＰ、１２．１％減に上方修正　１０−１２月期

　内閣府が１２日発表した０８年１０〜１２月期の国内総生 は、物価変動の影響を除いた……… (09:07) [記事全文]

- 金融サミット、気候変動も議論する可能性＝外交関 ター)
- 【株式・前引け】利益確定売りが先行、為替円高も ＴＯＰＩＸとも小幅反落 (3/12)（東洋経済）
- 『今回の上昇は本物か』【森田レポート】 (3/11)（今


asahi.com：朝日新聞社の速報ニュースサイト
Translated version of http://ww
http://translate.google.com/translate?prev=hp&hl=
Google
Google™
New Prius

This page was automatically translated from Japanese.
View original web page or mouse over text to view original lan

Business

Latest News

- The exchange of financial stocks fell slightly prominent lower

12 stocks in Tokyo, ahead of sell orders from the backlash of higher yesterday, with slightly lower values. Nikkei ... ... (11:13) [Full article]

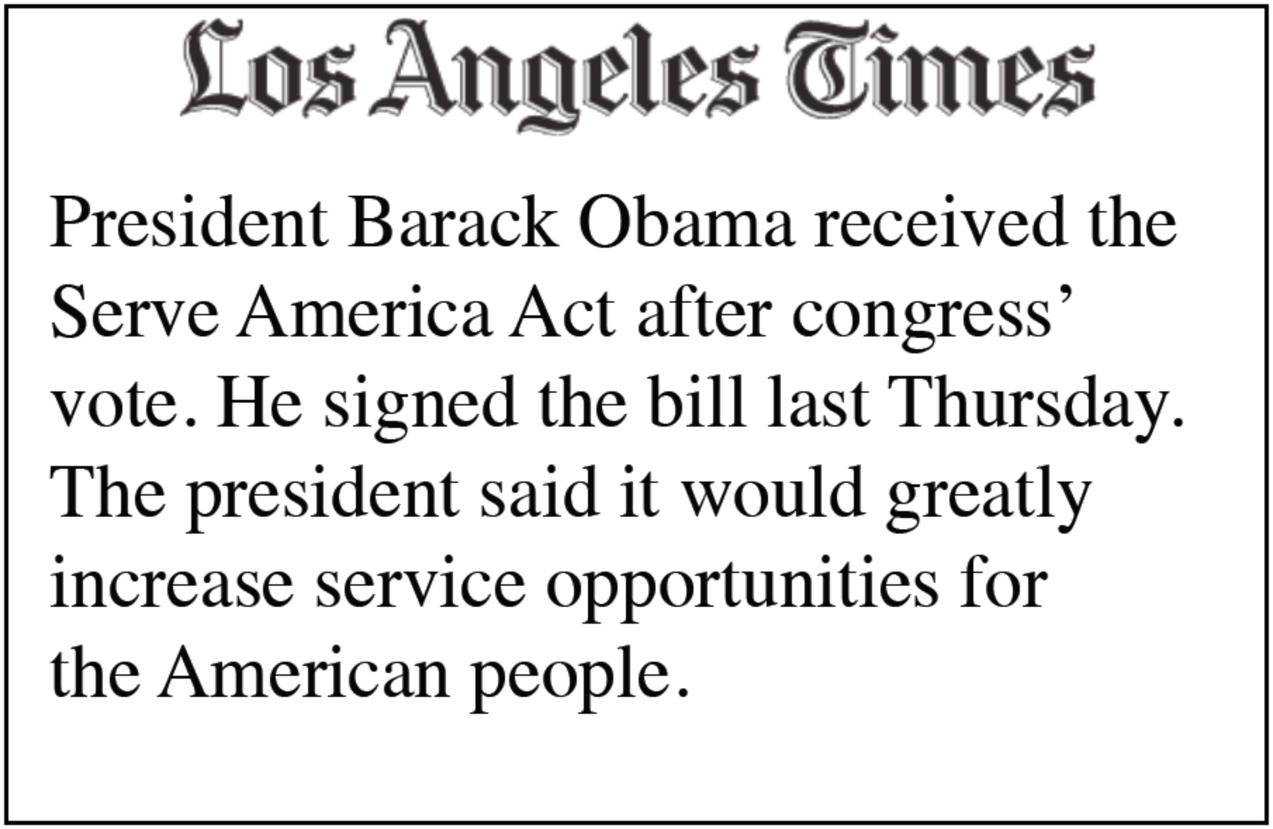
- Negotiation and integration of Japan Sompo Japan興亜to aggregate in three large camps

Sompo Japan Insurance and it's five to start the negotiations for the merger o NIPPONKOA Insurance Co., Ltd. No. 12, 2007, minutes ... ... (10:33) [Full

Deeper Understanding: Reference

Q: Who signed the Serve America Act?

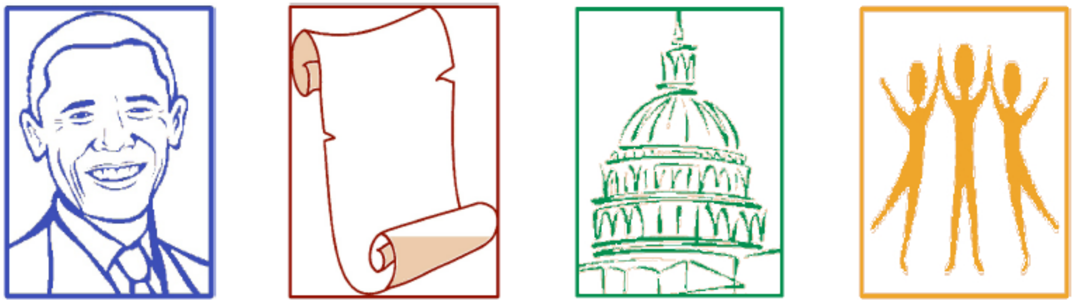
A: Barack Obama



Los Angeles Times

President Barack Obama received the Serve America Act after congress' vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Names vs. Entities



President Barack Obama received the Serve America Act after congress' vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Example Errors

Input

America Online announced on Monday that the company plans to update its instant messaging service.

Correct

America Online the company its

instant messaging service

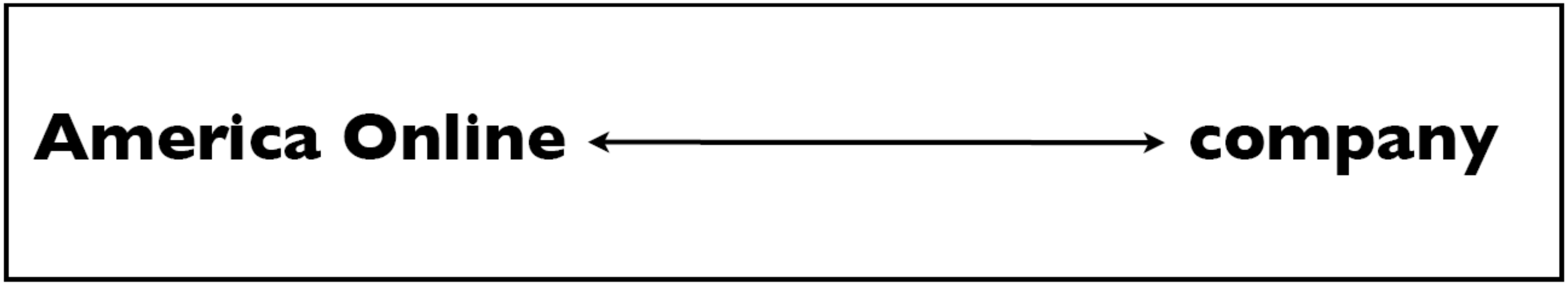
Guess

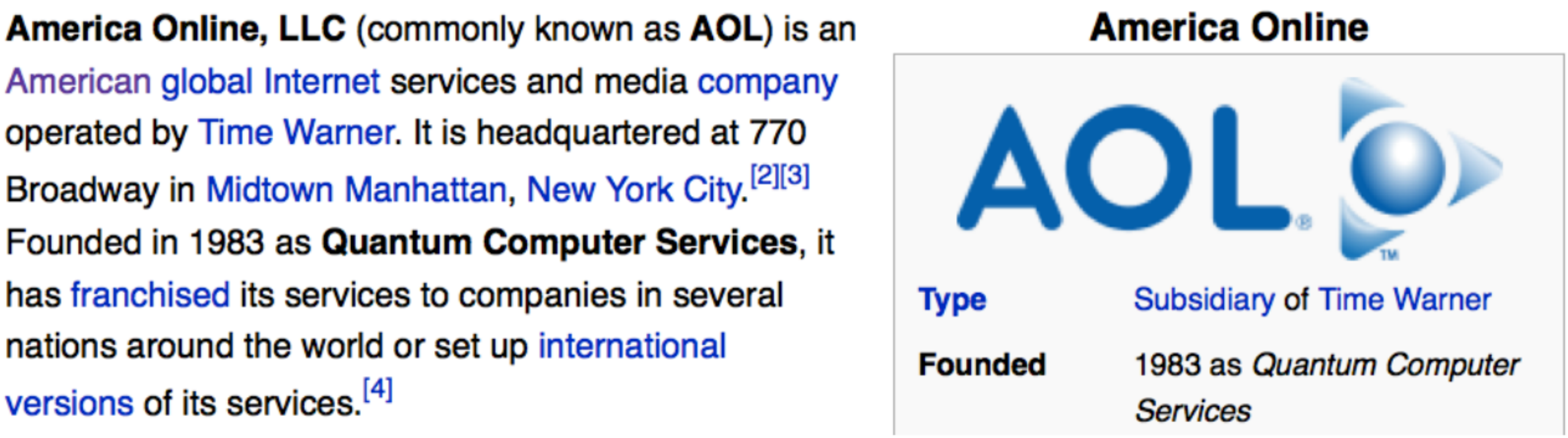
America Online

the company its

instant messaging service

Discovering Knowledge


America Online ←→ company


America Online, LLC (commonly known as **AOL**) is an American global Internet services and media company operated by Time Warner. It is headquartered at 770 Broadway in Midtown Manhattan, New York City.[2][3] Founded in 1983 as **Quantum Computer Services**, it has franchised its services to companies in several nations around the world or set up international versions of its services.[4]

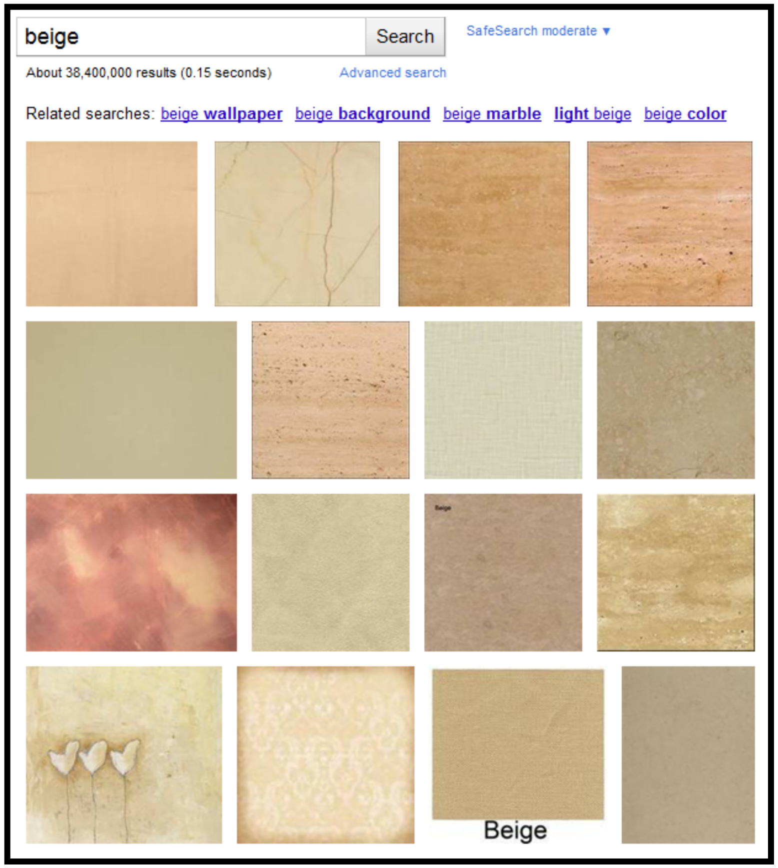
America Online

Type	Subsidiary of Time Warner
Founded	1983 as *Quantum Computer Services*

Grounded Language

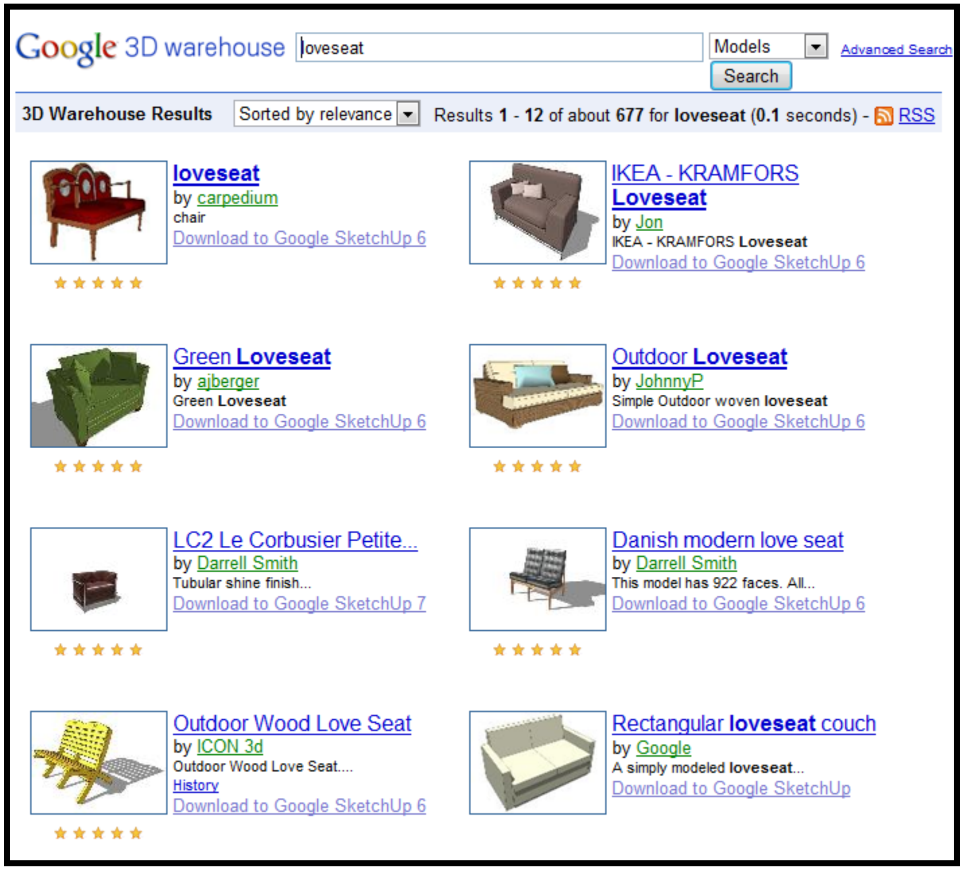
Grounding with Natural Data

… on the beige loveseat.



About 38,400,000 results (0.15 seconds)

Related searches: beige **wallpaper** beige **background** beige **marble** **light** beige beige **color**



3D Warehouse Results

Results **1** - **12** of about **677** for **loveseat** (**0.1** seconds) - RSS

loveseat
by carpedium
chair
Download to Google SketchUp 6

IKEA - KRAMFORS **Loveseat**
by Jon
IKEA - KRAMFORS **Loveseat**
Download to Google SketchUp 6

Green **Loveseat**
by ajberger
Green **Loveseat**
Download to Google SketchUp 6

Outdoor **Loveseat**
by JohnnyP
Simple Outdoor woven **loveseat**
Download to Google SketchUp 6

LC2 Le Corbusier Petite...
by Darrell Smith
Tubular shine finish...
Download to Google SketchUp 7

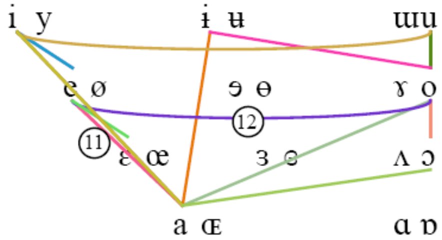
Danish modern love seat
by Darrell Smith
This model has 922 faces. All...
Download to Google SketchUp 6

Outdoor Wood Love Seat
by ICON 3d
Outdoor Wood Love Seat....
History
Download to Google SketchUp 6

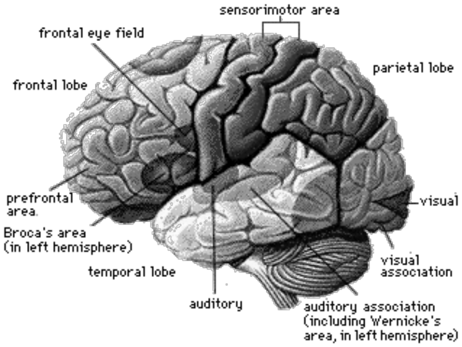
Rectangular **loveseat** couch
by Google
A simply modeled **loveseat**...
Download to Google SketchUp

What is Nearby NLP?

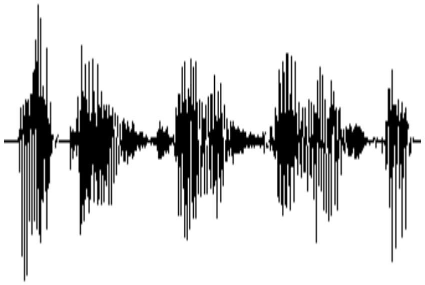
- Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it



- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP



What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice…

Class Requirements and Goals

- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models
    - Basic linguistics background
    - Strong coding skills (Java)
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps

- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

Some Early NLP History

- 1950's:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960's and 1970's: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - … but toy domains / grammars (SHRDLU, LUNAR)
- 1980's and 1990's: The Empirical Revolution
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*
- 2000+: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - *Begin to get both breadth and depth*

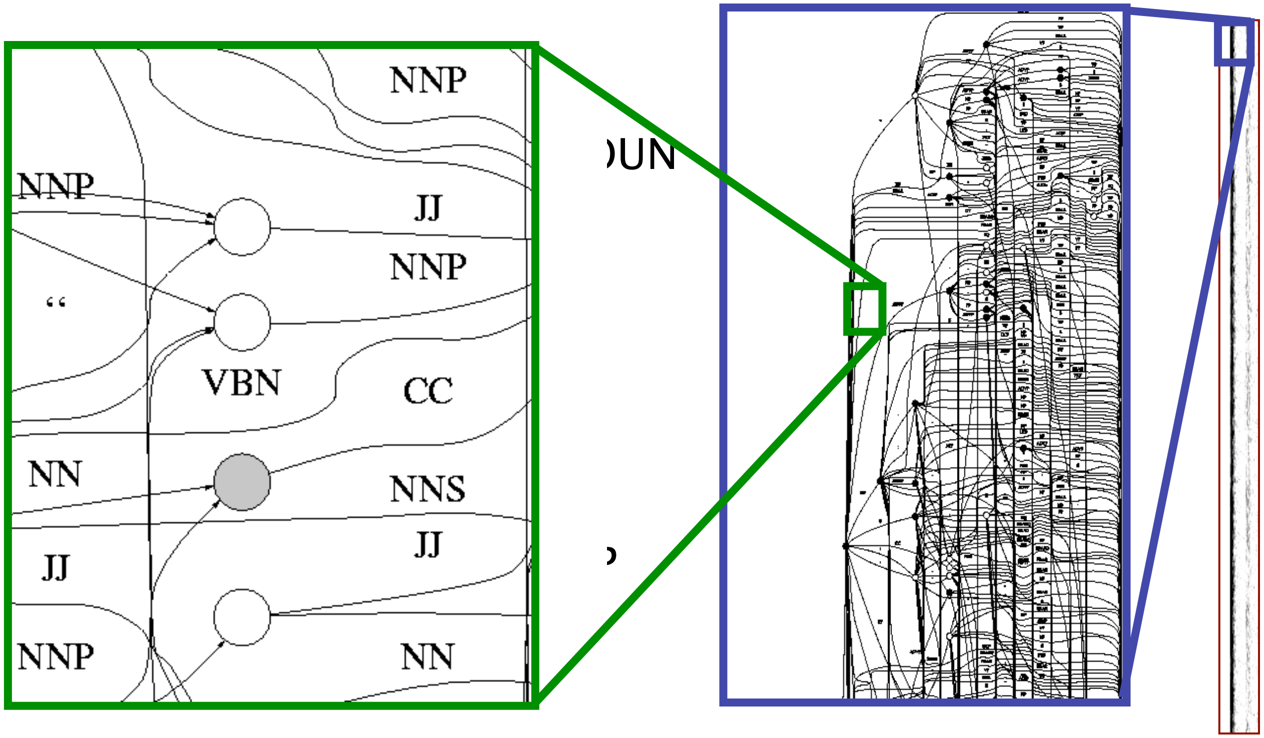
Problem: Structure

- Headlines:
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Why are these funny?

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



Classical NLP: Parsing

- Write symbolic or logical rules:

Grammar (CFG)

ROOT → S

S → NP VP

NP → DT NN

NP → NN NNS

NP → NP PP

VP → VBP NP

VP → VBP NP PP

PP → IN NP

Lexicon

NN → interest

NNS → raises

VBP → interest

VBZ → raises

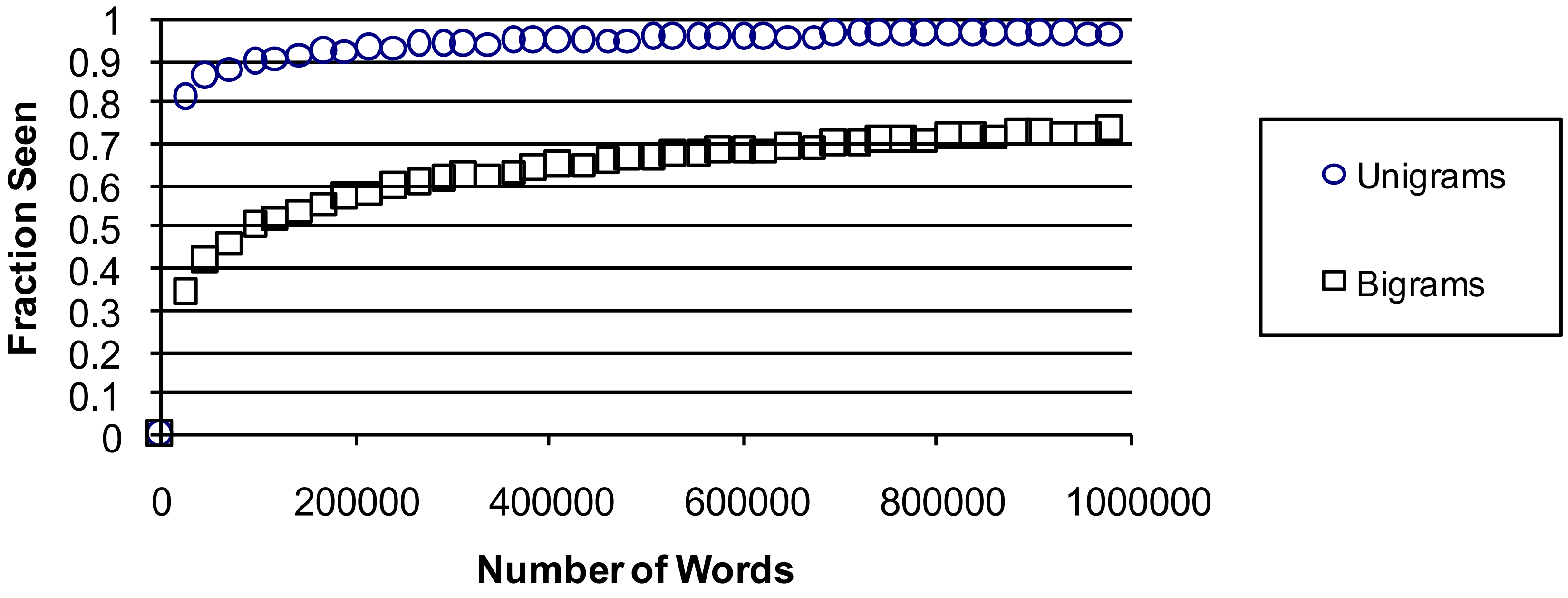
…

- Use deduction systems to prove parses from words
  - Minimal grammar on “Fed raises” sentence: 36 parses
  - Simple 10-rule grammar: 592 parses
  - Real-size grammar: many millions of parses

- This scaled very badly, didn’t yield broad coverage tools

Problem: Sparsity

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



Outline of Topics

- Words and Sequences
  - Speech recognition
  - N-gram models
  - Working with a lot of data
- Structured Classification
- Trees
  - Syntax and semantics
  - Syntactic MT
  - Question answering
- Machine Translation
- Other Topics
  - Reference resolution
  - Summarization
  - Diachronics
  - …

What’s Next?

- Next class: noisy-channel models and language modeling
  - Introduction to machine translation and speech recognition
  - Start with very simple models of language, work our way up
  - Some basic statistics concepts that will keep showing up
- If you don’t know what conditional probabilities and maximum likelihood estimators are, read up!
- Reading on the web

A Puzzle

- You have already seen N words of text, containing a bunch of different word types (some once, some twice…)
- What is the chance that the N+1st word is a new one?