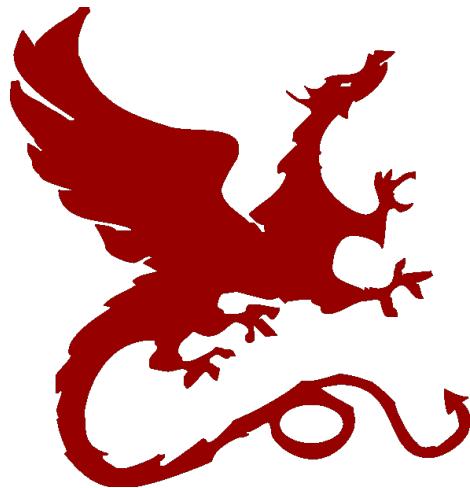


# Algorithms for NLP



## Lecture 1: Introduction

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley



# Course Information

[www.cs.cmu.edu/~tbergkir/11711fa17/](http://www.cs.cmu.edu/~tbergkir/11711fa17/)

## 11-711: Algorithms for NLP, Fall 2017

Instructors: [Taylor Berg-Kirkpatrick](#) and [Robert Frederking](#)

Lecture: Tuesday and Thursday 1:30pm-2:50pm, DH 1212

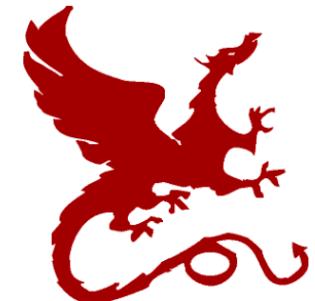
Recitation: Friday 1:30pm-2:20pm, MM 103

Office Hours: TBA

TAs: [Hieu Pham](#) and [Maria Ryskina](#)

Office Hours: TBA

Forum: Piazza (link to appear soon)



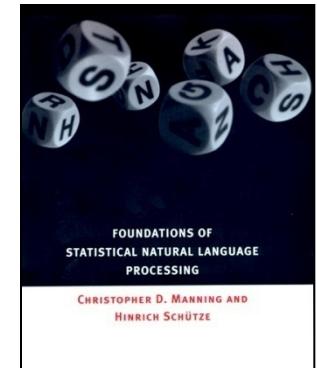
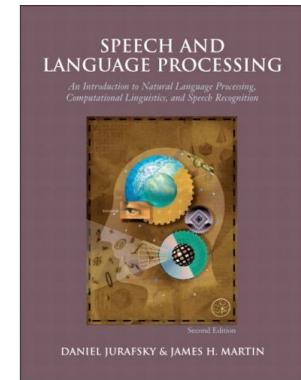
## Piazza setup soon!

The screenshot shows the Piazza interface. At the top, there's a navigation bar with 'piazza' and 'CS 288'. Below it is a toolbar with 'project1', 'project2', 'project3', 'project4', 'project5', 'New Post', and a search bar. On the left, there are sections for 'PINNED' and 'WEEK 8/10 - 8/16'. A yellow-highlighted post in the pinned section reads: '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 teaching staff...'. To the right, a note titled 'note' with 5 views is displayed, containing the same introductory text.



# Course Requirements

- **Prerequisites:**
  - Upper division algorithms (dynamic programming)
  - Mastery of basic probability
  - Strong skills in Java or equivalent
  - Deep interest in language
- **Work and Grading:**
  - Four assignments (individual, jars + write-ups)
- **Books:**
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2<sup>nd</sup> Edition (not 1<sup>st</sup>)
  - Also: Manning and Schuetze, Foundations of Statistical NLP





# 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?



# Language Technologies



## Goal: Deep Understanding

- Requires context, linguistic structure, meanings...

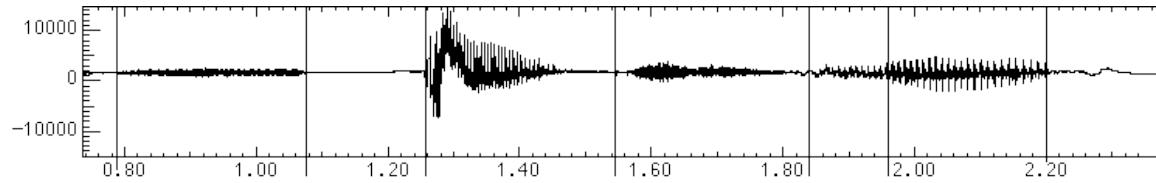
## Reality: Shallow Matching

- 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



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



# Text Data is Superficial

---

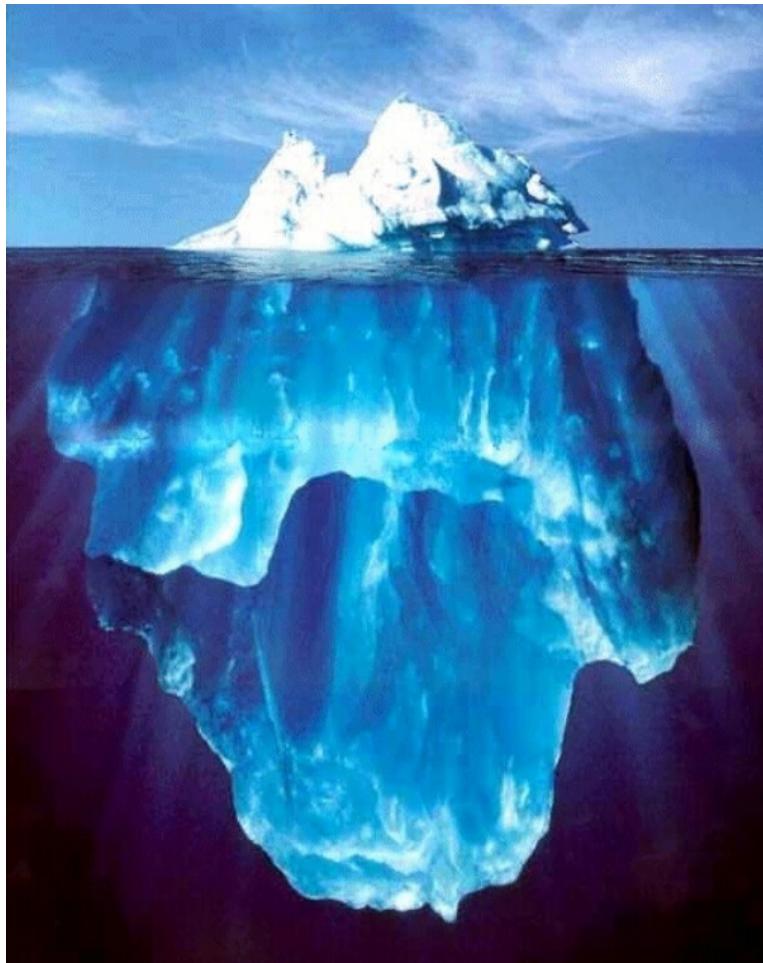
An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.





# ... But Language is Complex

---

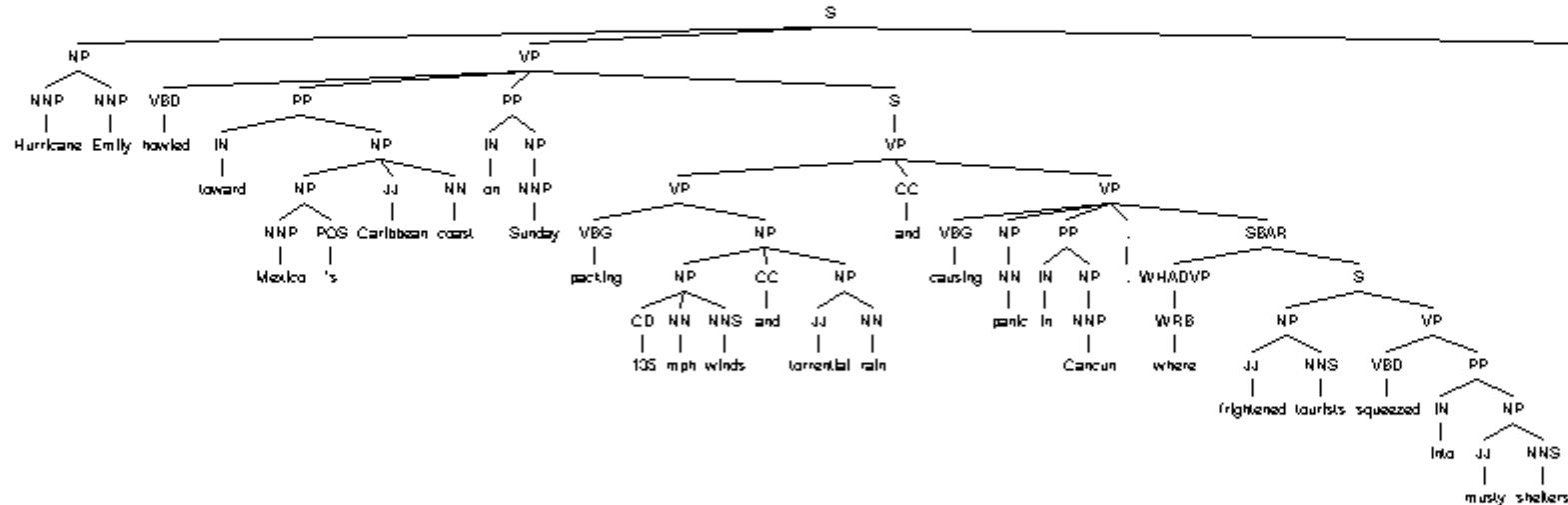


An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.

---



# Syntactic Analysis



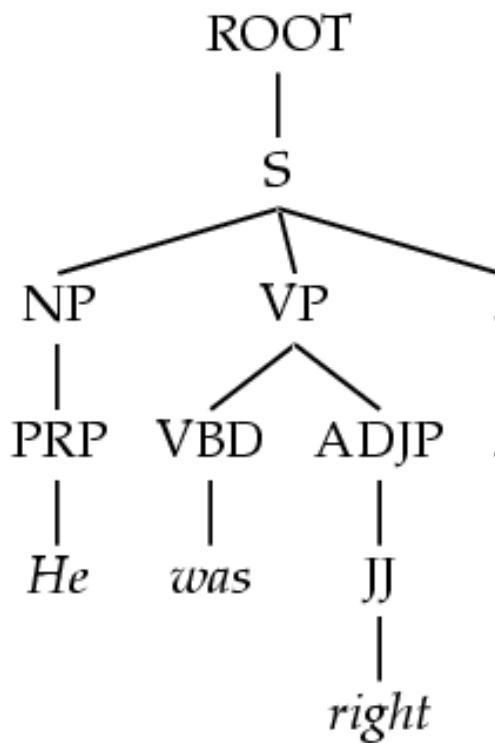
Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- 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

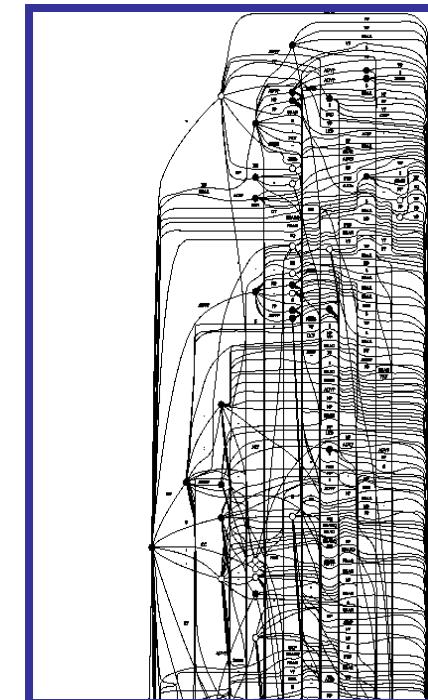
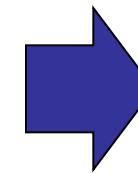


$\text{ROOT} \rightarrow \text{S}$

$\text{S} \rightarrow \text{NP VP} .$

$\text{NP} \rightarrow \text{PRP}$

$\text{VP} \rightarrow \text{VBD ADJ}$

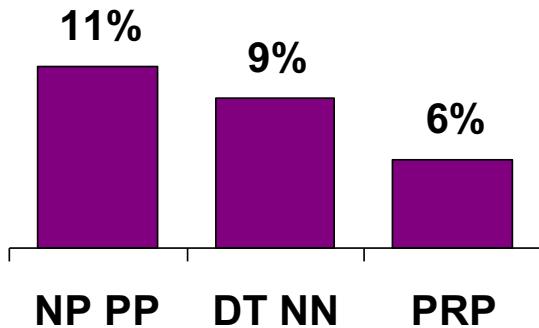




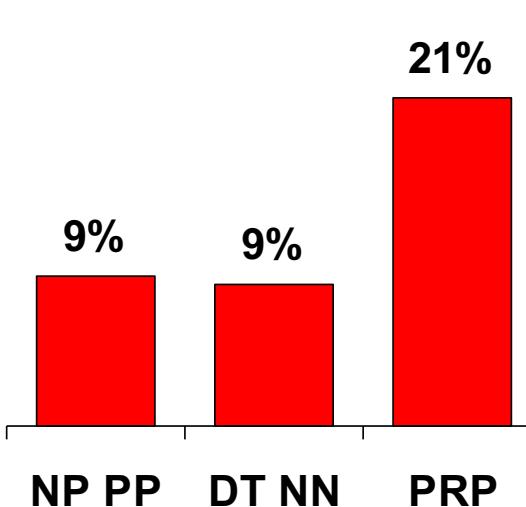
# Corpus-Based Methods

- It gives us statistical information

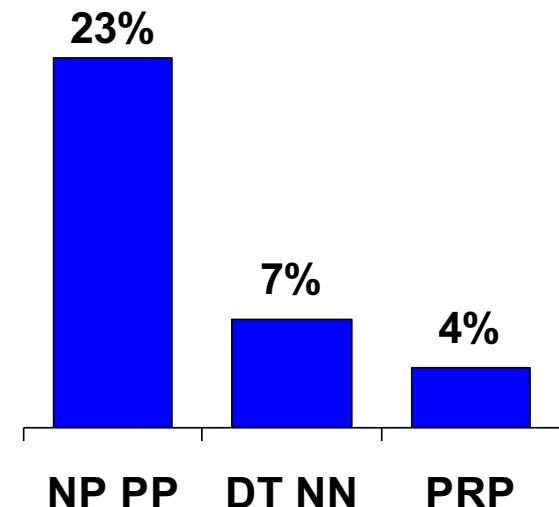
All NPs



NPs under S



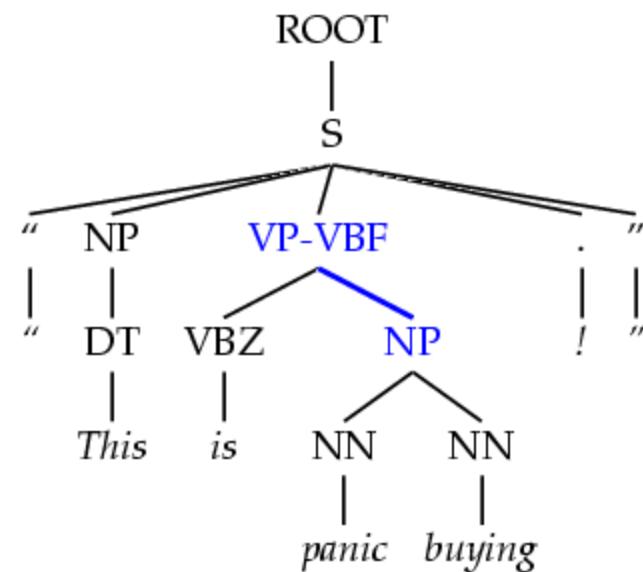
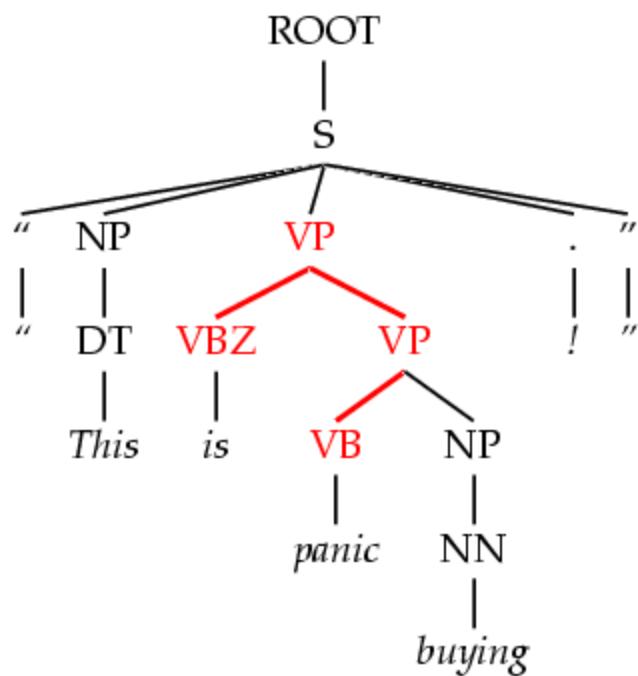
NPs under VP





# 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:
    - Segmentation

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

The screenshot shows a Google search results page. The search query is "any US states' capitals are also their largest cities?". The results are displayed under the "Web" tab. A message indicates that the search did not find any documents. Below the results, there are suggestions for improving the search query.

Google™ Web Images Groups News Froogle Local more »

any US states' capitals are also their largest cities? Search

Web

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.

Google Home -- Business Solutions - About Google

[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](http://www.answers.com/topic/capital-of-wyoming) - 21k - [Cached](#) - [Similar pages](#)

[Cheyenne: Weather and Much More From Answers.com](#)

Cheyenne ( shē-ān' , -ĕn' ) The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.  
[www.answers.com/topic/cheyenne-wyoming](http://www.answers.com/topic/cheyenne-wyoming) - 74k - [Cached](#) - [Similar pages](#)



# Example: Watson

"a camel is a horse designed by"

About a multilingual free encyclopedia Wiktionary [wɪkʃənri] n., a wiki-based Open Content dictionary Wiktionary

► Des One Vogu en.w a ca a ca analo Alter en.w Re: Re: A to: R www The Jan 4 com www A ca Sep 1 com bette Why Jun 2 variat www If a camel is a horse de If a camel is a horse design

Entry Discussion Read Edit History Search

a camel is a horse designed by a committee

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

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

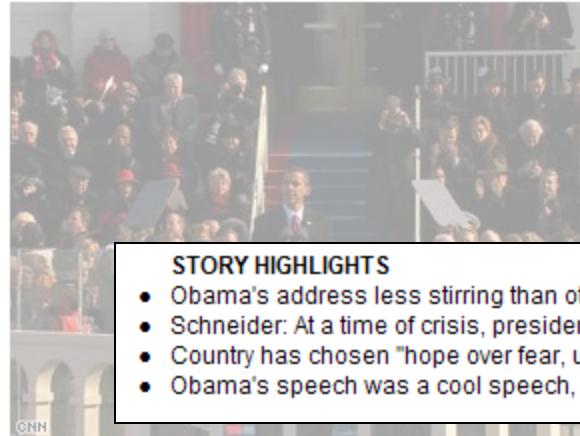
- [Re: A camel is a horse designed by committee SR 16/April/04](#)
  - [Re: A camel is a horse designed by committee Henry 18/April/04](#)



# 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.



## 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

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 begin

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.

**WAtoday.com.au**

**LIFE & STYLE**

**THE DAILY CALLER**

**Judge tells Lohan she's no star in his courtroom**

**Lindsay Lohan charged with theft**

**Join the conversation**

**Top Life & Style articles**

**Story Tools**

**WA Today Jobs**

**WAtoday.com.au**

**LIFE & STYLE**

**Choose Qantas & enjoy: Genuine Aussie Hospitality Complimentary food & wine More than 400 on-demand movies & TV shows Available between the mainland US and Australia**

**You are here: Home > Life & Style > People > Article**

**Lindsay Lohan charged with theft**

**February 10, 2011**

**Join the conversation**

**Top Life & Style articles**

**Story Tools**

**WA Today Jobs**

**THE DAILY CALLER**

**Log In | Register**

**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.**



# 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

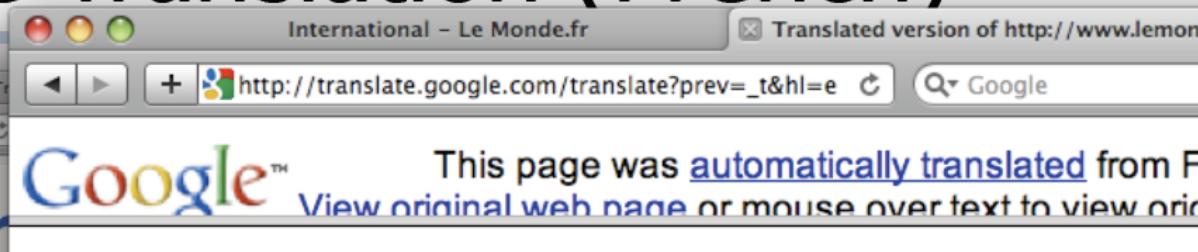
**Video** Anniversary of the Tibetan rebellion: China on guard



- 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)

International – Le Monde.fr      Translated version of <http://www.lemonde.fr>

**Le Monde.fr**  
Mise à jour à 05h17 - Paris

**"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

[Portfolio](#) | [Reportage](#) | [Vidéo](#)

**Accord sur la TVA : "Sarkozy gagne le pire moment"**

Les ministres des finances européens ont trouvé un compromis autorisant la réduction de certains secteurs, dont la restauration.

**Compte rendu** Réactions mitigées à une baisse de la TVA

**Les faits** Les taux réduits de TVA au

**Face aux déficits, la hausse paraît inéluctable**

Le gouvernement exclut une augmentation de la TVA. Philippe Séguin tire la sonnette d'alarme.

**Infographie** Finances publiques : les prévisions gouvernementales

**Les faits** La crise avive le débat fiscal

[Eclairage](#) | [Compte rendu](#)

**This page was automatically translated from French**  
[View original web page](#) or mouse over text to view original

## "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 the 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 on reduction in VAT

The deal to cut VAT rates in some sectors



# 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)

asahi.com (朝日新聞社) : ビジ... Translated

トップ ニュース スポーツ エンタメ ライフ  
社会 ビジネス 政治 國際 文化 サイエンス  
現在位置: asahi.com > ニュース > ビジネス

トップ ニュース 為替 株式 金利 トピックス 投  
東洋経済ニュース ロイターニュース 宝くじ CSR

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=外交閣  
タード

▶ 【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://www.asahi.com/business/

http://translate.google.com/translate?prev=hp&hl=

Google

Google™ This page was automatically translated from Japanese.  
[View original web page](#) or mouse over text to view original language.

## 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 of NIPPONKOA Insurance Co., Ltd. No. 12, 2007, minutes ..... (10:33) [Full article]



New Prius



## Deeper Understanding: Reference

---

Q: Who signed the Serve America Act?

A: Barack Obama

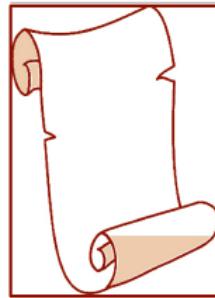
### Los Angeles Times

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.<sup>[2][3]</sup> 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.<sup>[4]</sup>

**America Online**



**Type**

Subsidiary of Time Warner

**Founded**

1983 as *Quantum Computer Services*



# Grounded Language





# Grounding with Natural Data

*... on the beige loveseat.*

beige

Search SafeSearch moderate ▾

About 38,400,000 results (0.15 seconds) Advanced search

Related searches: [beige wallpaper](#) [beige background](#) [beige marble](#) [light beige](#) [beige color](#)

Beige

Google 3D warehouse loveseat Models Advanced Search

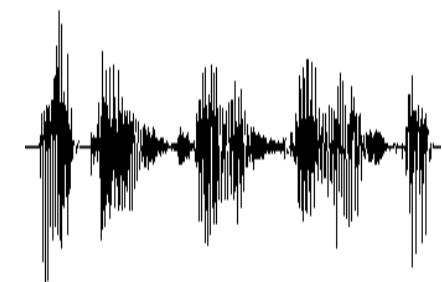
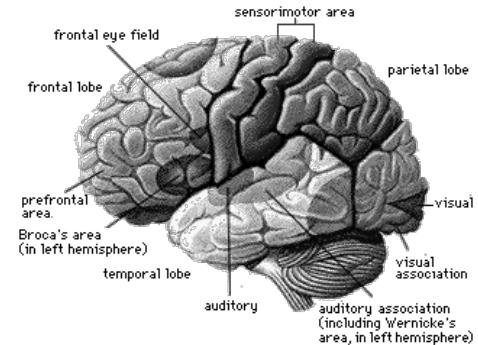
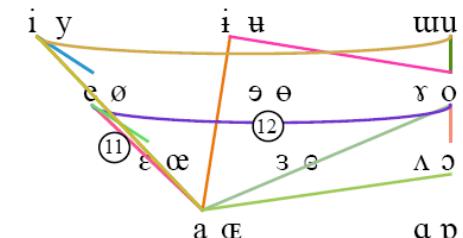
3D Warehouse Results Sorted by relevance Results 1 - 12 of about 677 for loveseat (0.1 seconds) - RSS

 loveseat by <a href="#">carpedium</a> chair <a href="#">Download to Google SketchUp 6</a> ★★★★★	 IKEA - KRAMFORS Loveseat by <a href="#">Jon</a> IKEA - KRAMFORS Loveseat <a href="#">Download to Google SketchUp 6</a> ★★★★★
 Green Loveseat by <a href="#">ajberger</a> Green Loveseat <a href="#">Download to Google SketchUp 6</a> ★★★★★	 Outdoor Loveseat by <a href="#">JohnnyP</a> Simple Outdoor woven loveseat <a href="#">Download to Google SketchUp 6</a> ★★★★★
 LC2 Le Corbusier Petite... by <a href="#">Darrell Smith</a> Tubular shine finish... <a href="#">Download to Google SketchUp 7</a> ★★★★★	 Danish modern love seat by <a href="#">Darrell Smith</a> This model has 922 faces. All... <a href="#">Download to Google SketchUp 6</a> ★★★★★
 Outdoor Wood Love Seat by <a href="#">ICON 3d</a> Outdoor Wood Love Seat.... <a href="#">History</a> <a href="#">Download to Google SketchUp 6</a> ★★★★★	 Rectangular loveseat couch by <a href="#">Google</a> A simply modeled loveseat... <a href="#">Download to Google SketchUp</a>



# 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

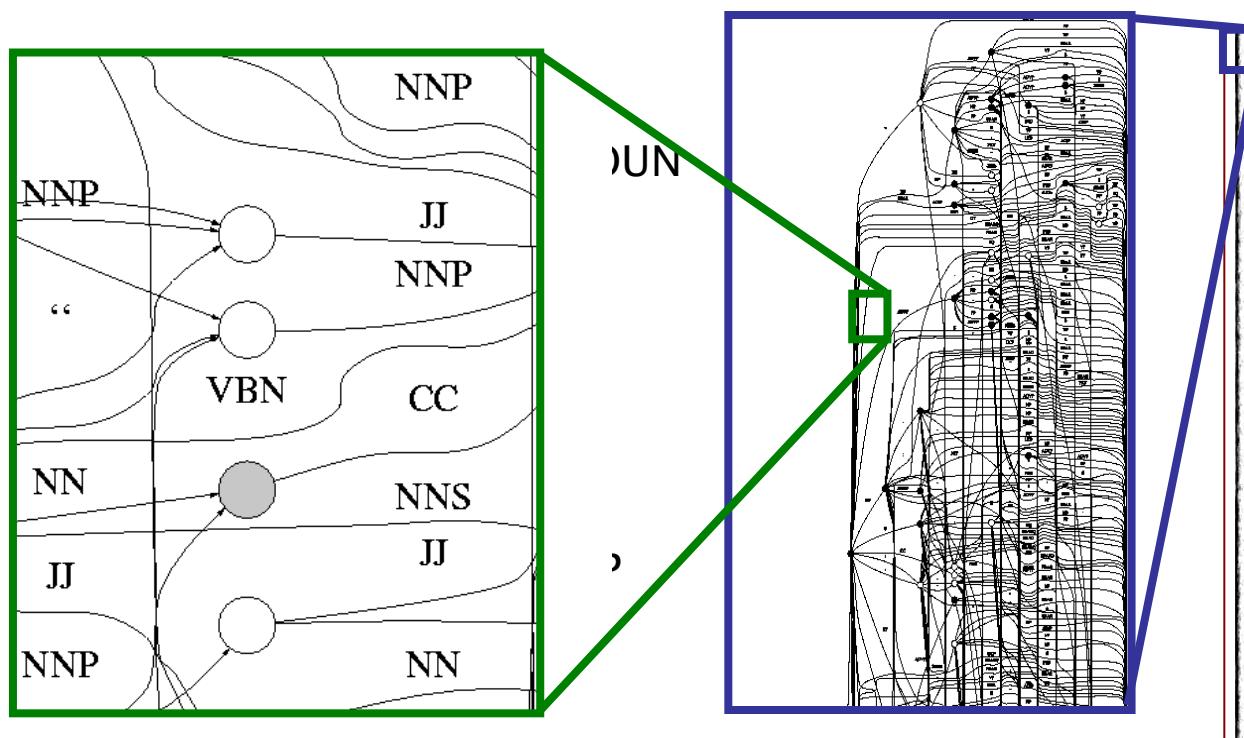
---

- 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)

$\text{ROOT} \rightarrow S$

$S \rightarrow NP\ VP$

$NP \rightarrow DT\ NN$

$NP \rightarrow NN\ NNS$

$NP \rightarrow NP\ PP$

$VP \rightarrow VBP\ NP$

$VP \rightarrow VBP\ NP\ PP$

$PP \rightarrow IN\ NP$

Lexicon

$NN \rightarrow \text{interest}$

$NNS \rightarrow \text{raises}$

$VBP \rightarrow \text{interest}$

$VBZ \rightarrow \text{raises}$

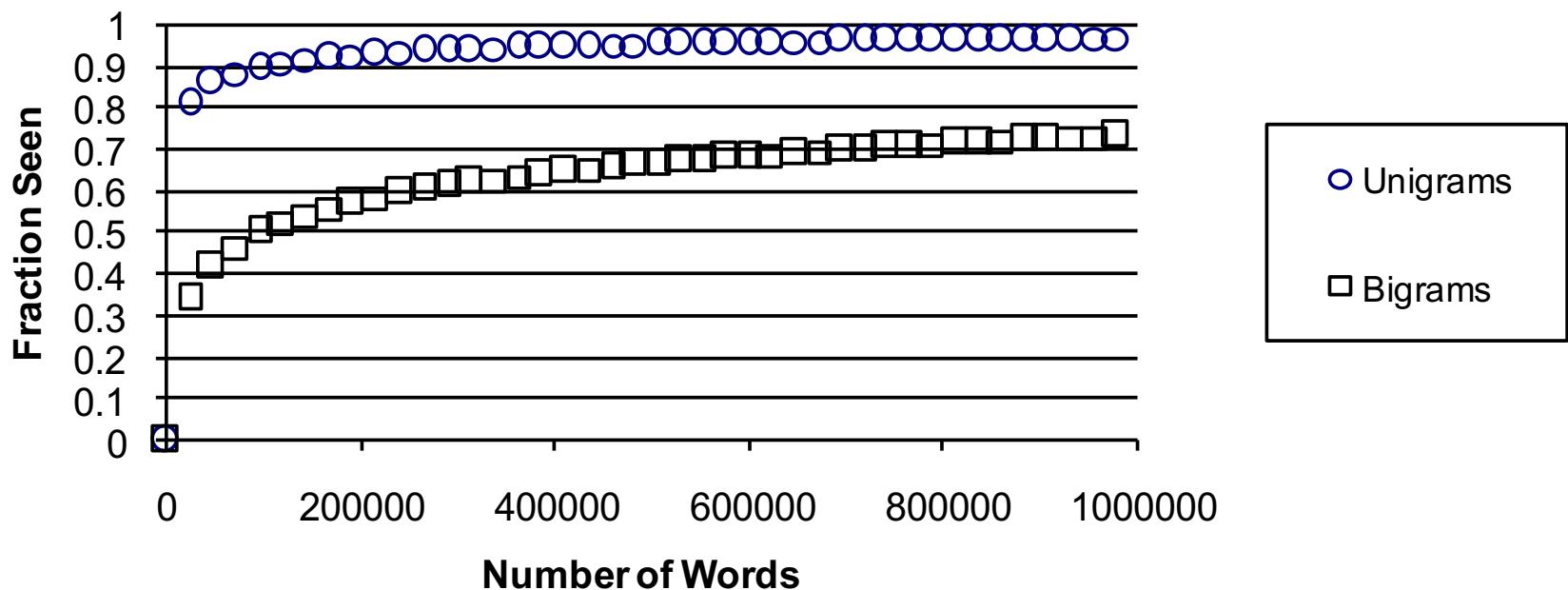
...

- 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



---

# AI: Where Do We Stand?

Hollywood

R2D2



KITT



Wall-E



'80

Rule based  
approaches



'90

Early statistical  
approaches



'00

Modern statistical approaches  
Google Translate '08



Reality



# Example Translation

---

## Près de la moitié des utilisateurs de Google News ne lisent que les titres (sondage)

Près de la moitié des utilisateurs du portail d'actualités Google News se contentent de lire les titres, sans cliquer dessus pour accéder aux articles auxquels ils renvoient sur les sites des différents journaux, selon un sondage publié mardi.

"Bien que Google génère une certaine fréquentation sur les (sites des) journaux, il en détourne aussi une bonne partie", note Ken Doctor, analyste au cabinet de conseil Outsell qui a réalisé ce sondage, soulignant que "pas moins de 44% des visiteurs de Google News regardent les titres sans aller sur les sites des journaux".

Ce sondage semble apporter de l'eau au moulin des critiques de Google, comme le magnat des médias Rupert Murdoch, patron de News Corp., qui accuse Google et d'autres agrégateurs de ne pas partager les recettes publicitaires avec les sites vers lesquels ils proposent des liens.

## Almost half of Google News users to read only the titles (poll)

Almost half of the portal users to Google News News merely read the headlines without clicking to access the items to which they refer to the websites of various newspapers, according to a poll released Tuesday.

"Although Google generates some traffic on the (sites) newspapers, it also diverts much," said Ken Doctor, an analyst at consulting firm Outsell, which conducted the survey, noting that "not less than 44% of Google News visitors looking titles without going on newspaper sites.

This survey seems to bring water to appease critics of Google, such as media mogul Rupert Murdoch, News Corp. boss., Accusing Google and other aggregators do not share advertising revenue with the sites to which they provide links.



---



# Next Week

---

- We're going to do an experiment in competitive parsing
- Polls:
  - How many have an EECS research account?
  - How many have a laptop they can bring next week?
  - How many know what a prepositional phrase is?
- Also: Assignment 1 will be out very soon, and assignments will move fast



# Machine Translation

---

## Original Text

新华网石家庄11月16日电（记者 张涛）11月15日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至1℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

## Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 °C. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.



# Discourse Ambiguity: Reference

---

“The American Medical Association voted yesterday to install the heir apparent as its president-elect, rejecting a strong, upstart challenge by a District doctor who argued that the nation’s largest physicians’ group needs stronger ethics and new leadership.”

- Can link repeated mentions of an entity both inside a document and across documents, model variation in spelling, transliteration, etc.
- SOTA: varies widely, 70-90%



# NLP: Annotation

---

- Much of NLP is annotating text with structure which specifies how it's assembled.
  - Syntax: grammatical structure
  - Semantics: “meaning,” either lexical or compositional

*John      bought      a      blue      car*



# Why is NLP Hard?

---

- The core problems:
  - Ambiguity
  - Sparsity
  - Scale
  - Unmodeled Variables



# Syntactic Ambiguities

---

- Maybe we're sunk on funny headlines, but normal, boring sentences are unambiguous?

*Fed raises interest rates 0.5 % in a measure against inflation*



# Disambiguation for Applications

---

- Sometimes life is easy
  - Can do text classification pretty well just knowing the set of words used in the document, same for authorship attribution
  - Word-sense disambiguation not usually needed for web search because of majority effects or intersection effects (“jaguar habitat” isn’t the car)
- Sometimes only certain ambiguities are relevant

*he hoped to record a world record*

- Other times, all levels can be relevant (e.g., translation)



# Language isn't Adversarial

---

- One nice thing: we know NLP can be done!
- Language isn't adversarial:
  - It's produced with the intent of being understood
  - With some understanding of language, you can often tell what knowledge sources are relevant
- But most variables go unmodeled
  - Some knowledge sources aren't easily available (real-world knowledge, complex models of other people's plans)
  - Some kinds of features are beyond our technical ability to model (especially cross-sentence correlations)



**RULE 15:**  
S(x0:NP, x1:VP, x2:PUNC)  
 $\rightarrow x_0, x_1, x_2$

“These 7 people include astronauts coming from France and Russia”

**RULE 14:**  
VP(x0:VBP, x1:NP)  
 $\rightarrow x_0, x_1$

“include astronauts coming from France and Russia”

**RULE 16:**  
NP(x0:NP, x1:VP)  
 $\rightarrow x_1, \text{的}, x_0$

“astronauts coming from France and Russia”

**RULE 11:**  
VP(VBG(coming), PP(IN(from), x0:NP))  
 $\rightarrow \text{来自}, x_0$

“coming from France and Russia”

**RULE 13:**  
NP(x0>NNP, x1:CC, x2>NNP)  
 $\rightarrow x_0, x_1, x_2$

“France and Russia”

**RULE 10:**  
NP(x0:DT, CD(7), NNS(people))  
 $\rightarrow x_0, 7\text{人}$

“these 7 people”

**RULE 1:**  
DT(these)  
 $\rightarrow \text{这}$

“these”

**RULE 2:**  
VBP(include)  
 $\rightarrow \text{中包括}$

“include”

**RULE 4:**  
NNP(France)  
 $\rightarrow \text{法国}$

“France”

**RULE 5:**  
CC(and)  
 $\rightarrow \text{和}$

“&”

**RULE 6:**  
NNP(Russia)  
 $\rightarrow \text{俄罗斯}$

“Russia”

**RULE 8:**  
NP(NNS(astronauts))  
 $\rightarrow \text{宇航员}$

“astronauts”

**RULE 9:**  
PUNC(.)  
 $\rightarrow .$

“.”

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Derivation Tree



# Machine Translation

---

## **Atlanta, preso il killer del palazzo di Giustizia**

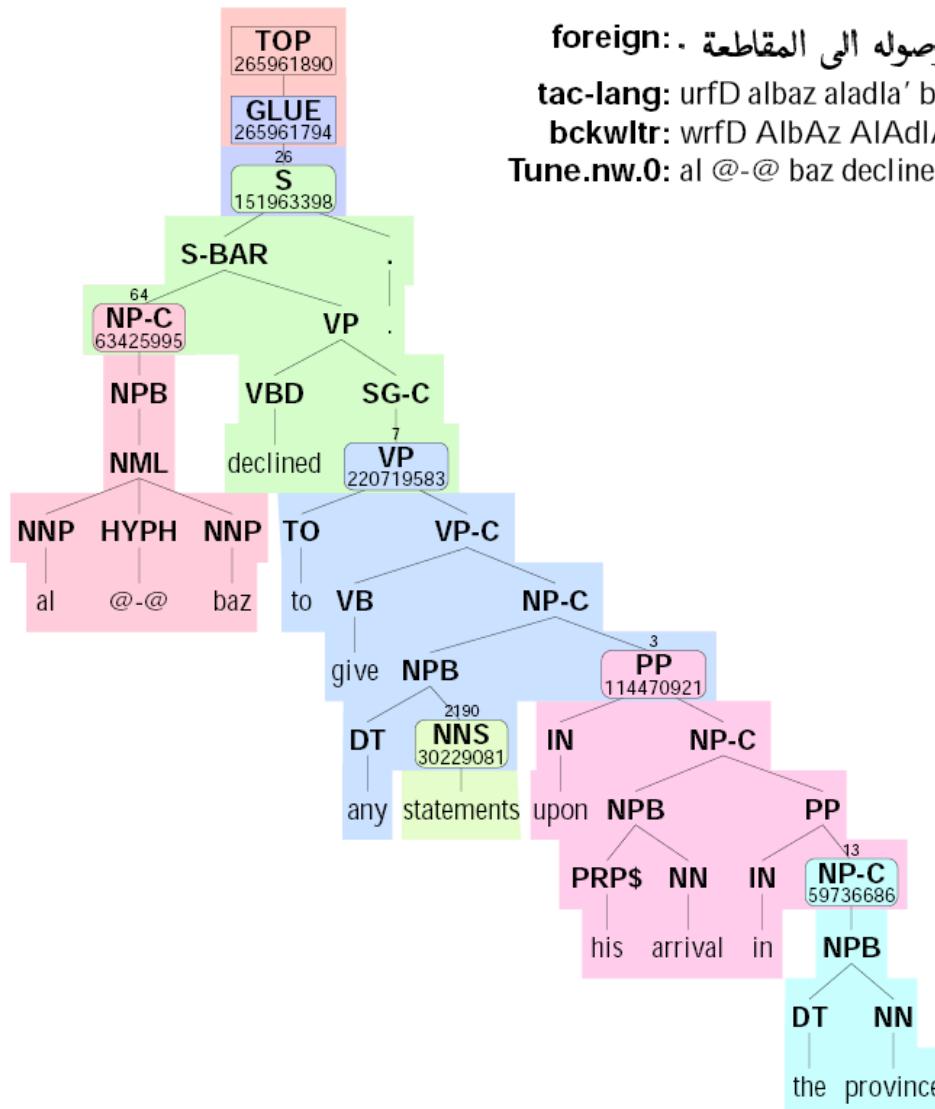
**ATLANTA** - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

## **Atlanta, taken the killer of the palace of Justice**

**ATLANTA** - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

- Translation systems encode:
  - Something about fluent language (next class)
  - Something about how two languages correspond (middle of term)
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators

# Example Syntax-Based Translation



ورفض البار الادلاء باى تصريحات فور وصوله الى المقاطعة .

**tac-lang:** urfD albaz aladla' baá tSryHat fur uSulh alá almqaT'e .

**bckwltr:** wrfD AlbAz AlAdIA' bAY tSryHAt fwr wSwlh AIY AlmqATEp .

**Tune.nw.0:** al @-@ baz declined to make any statements upon his arrival in the province .



# The Dream

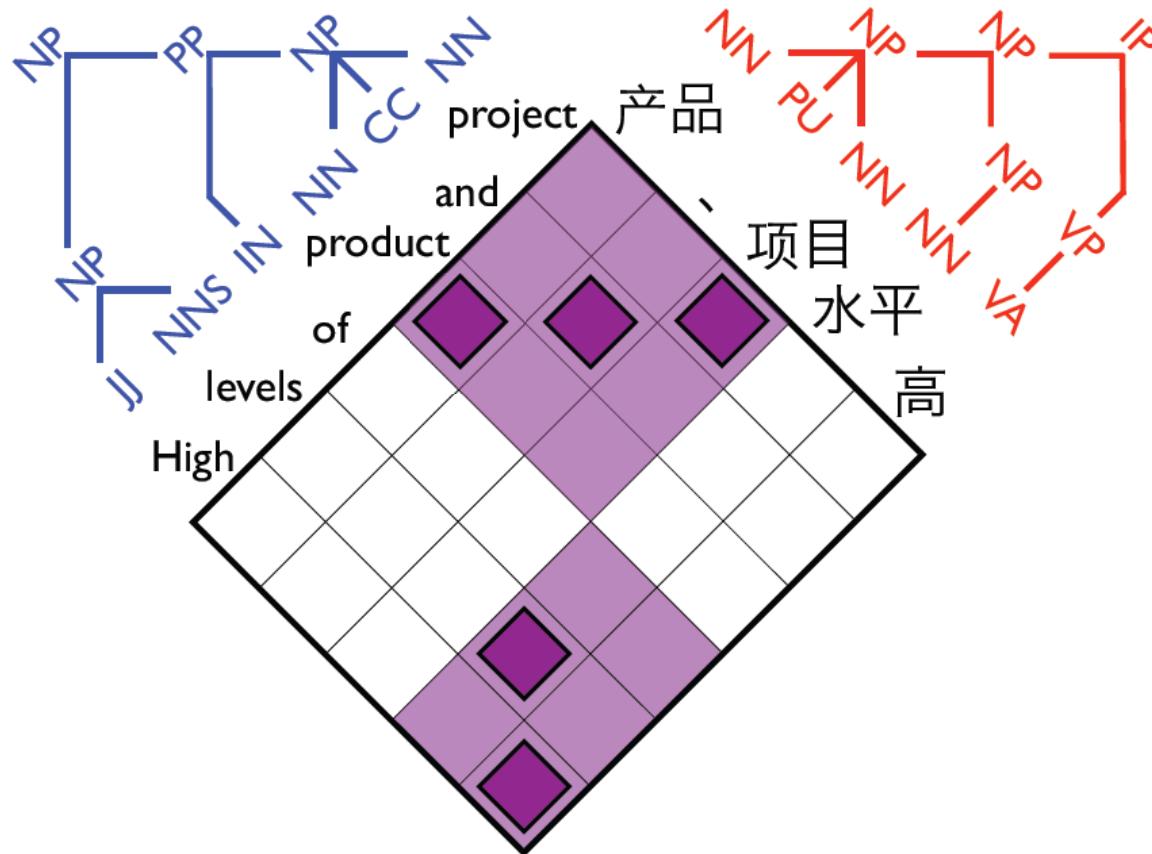
---

- It'd be great if machines could
  - Manage information in our email
  - Translate languages accurately
  - Help us process, summarize, and aggregate text information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge
- So:





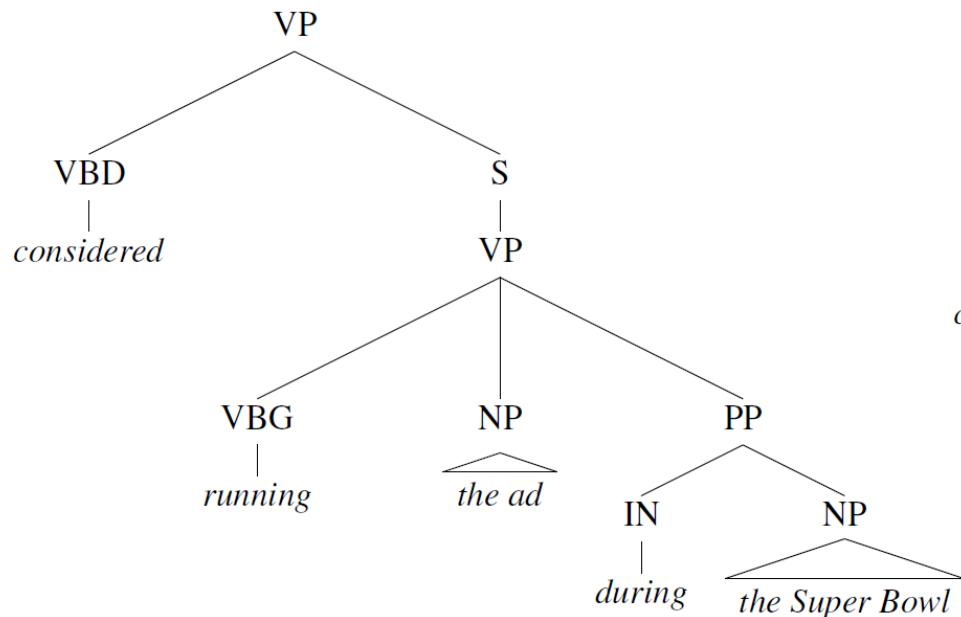
# Analysis and Alignment





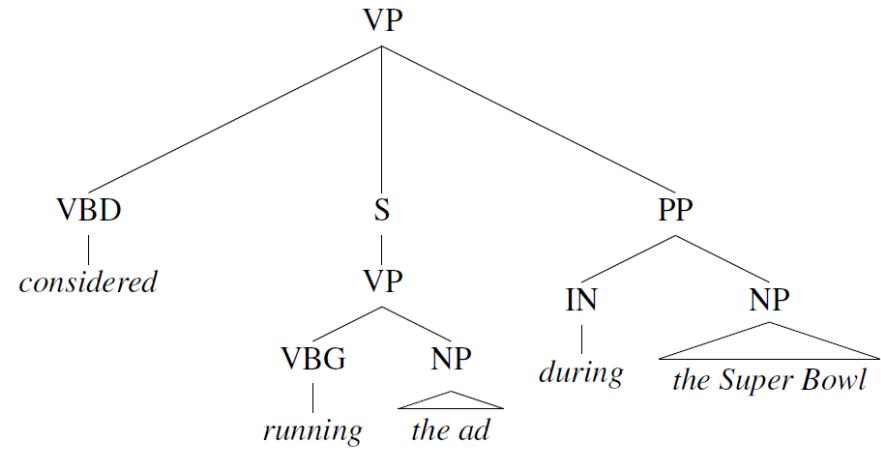
# Data and Knowledge: Parsing

*They considered running the ad during the Super Bowl.*



*running \* during: 3k*

*running it during: 239*

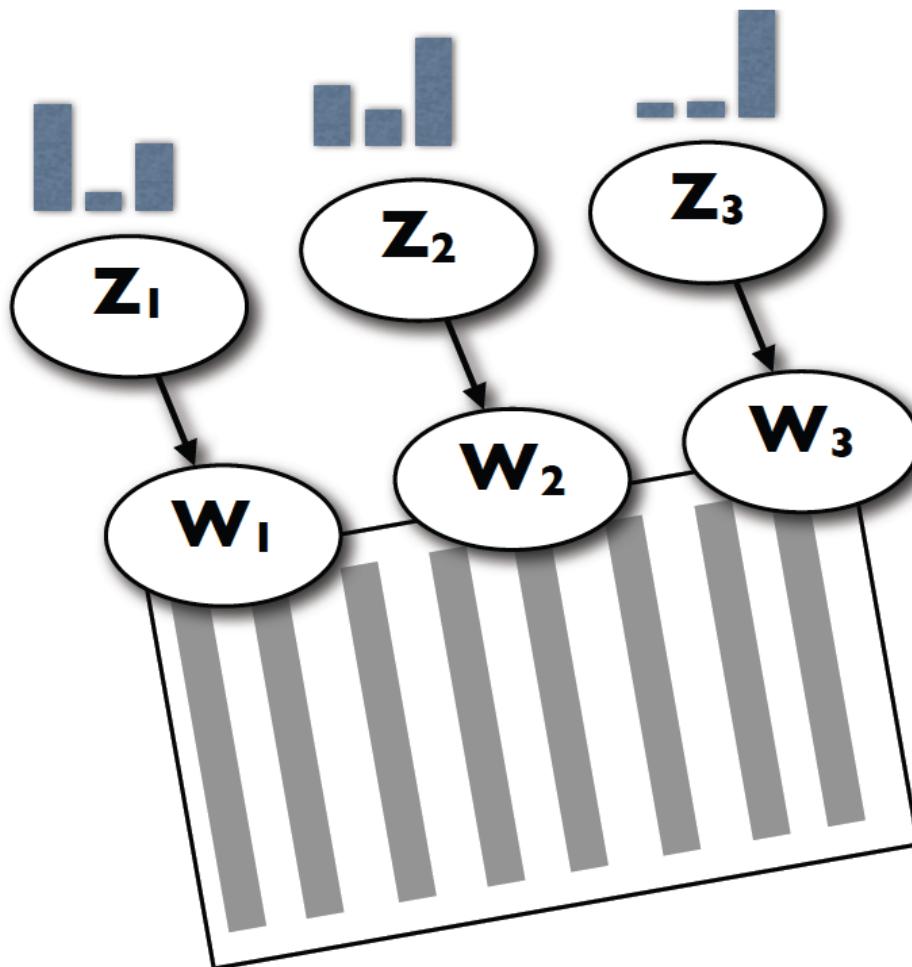


*considered \* during: 2k*

*considered it during: 112*



# Unsupervised Learning

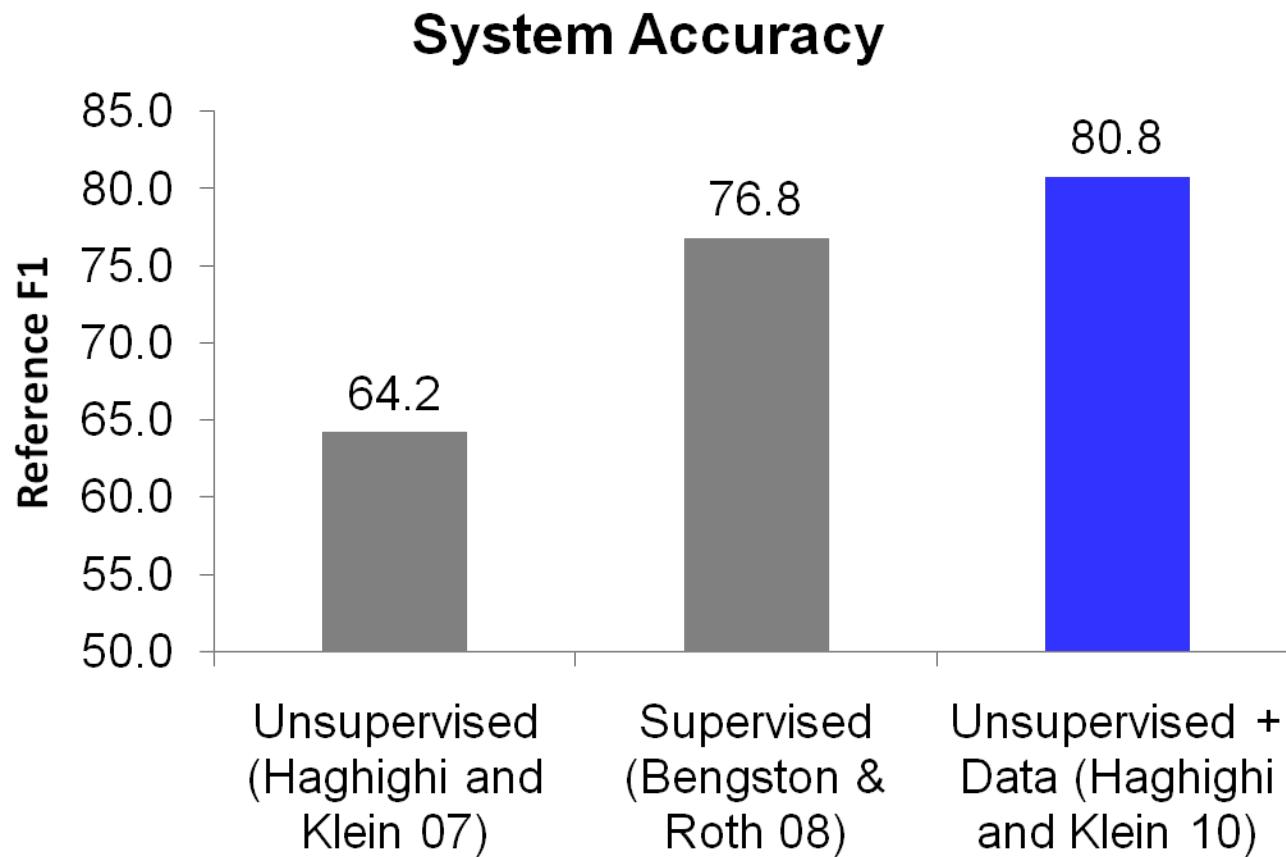


Parameters

0.20	0.45	0.35
0.55	0.28	0.17
0.31	0.51	0.18
0.06	0.70	0.24
0.62	0.13	0.15



# Coreference Systems





# Cross-Document Identity

**Dan Klein**  
Associate Professor  
Computer Science Division  
University of California at Berkeley

**Contact Information**

Email: klein@cs.berkeley.edu  
Mail: Dan Klein, Soda Hall, Bldg. 94720-1776  
Phone: (510) 643-0805 (email w/ "dklein")

**Research**

My research focuses on the automatic organization of natural language in some topics of interest to me are:

- Unsupervised language acquisition
- Machine translation
- Efficient algorithms for NLP
- Information extraction
- Linguistically rich models of language
- Integrating symbolic and statistical methods for NLP
- Organization of the web

[My group's web page \(the Berkeley Language Processing Group\)](#)

Our agent, the [Overmind](#), won the [StarCraft AI competition!](#)

**Background**

**Daniel Klein**  
[dklein@gmu.edu](mailto:dklein@gmu.edu)

I am Professor of Economics at George Mason University. I have a Ph.D. from George Mason University and New York University, where in both cases I studied the classical liberal traditions of economics. My research focuses on economic principles, public policy issues, and the tradition of Adam Smith and Friedrich Hayek.

I've published research on policy issues including toll roads, environmental auto emission, credit reporting, and the Food and Drug Administration. I've also written on spontaneous order, coordination, the discipline of opportunity, the demand and supply of assurance, why government officials believe in the goodness of bad policy, why people favor government intervention more than they should, and the relation between liberty, dignity, and responsibility.

Lately my research dwells on Adam Smith. Russ Roberts and I co-hosted a multipart audio book club on *The Theory of Moral Sentiments*.

**UCLA BRUINS** #1 IN NCAAC CHAMPION THE OFFICIAL WEBSITE OF UCLA ATHLETICS

SPORTS DONATE TICKETS RECRUITS FAN CENTRAL CAMPS ABOUT UCI

ROSTER | SCHEDULE/RESULTS | STATISTICS | NEWS | ARCHIVES

Like Sign Up to see what your friends like. Share Print Email Text RSS

**23 Dan Klein**

 **PROFILE**

Class: RS Sophomore	Height / Weight: 6-3 / 190
Hometown: Los Alamitos, Calif.	Position: RHP
High School: Servite	Bats/Throws: R/R

**Notes**  
Helped lead UCLA to two NCAA Regionals (2008, 2010) and the Bruins' third-ever trip to the College World Series in 2010...had a sensational season as the Bruins' closer in 2010...as a redshirt sophomore in 2010, recorded 10 saves, the second-highest single-season total in school history...selected in the third round (85th overall) of the 2010 MLB Draft by the Baltimore Orioles.

**2010**  
Made 39 appearances (all in relief), tying UCLA's single-season appearances record...also tied for the Pac-10 lead in appearances with Arizona State's Mitchell Lambson...served as UCLA's closer, notching 10 saves, the second-highest total single-season total in program history and the third-most saves among Pac-10 pitchers in 2010...earned All-Pac-10 Team honors, going 6-1 with a team-leading 1.90 ERA...tallied 55 strikeouts and 11 walks in 52.0 innings, limiting the opposition to



# Information Extraction

---

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

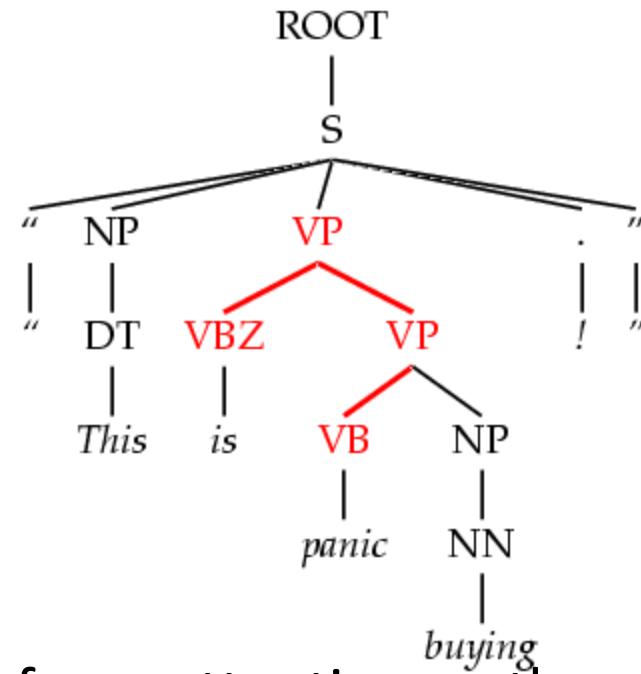
- SOTA: perhaps 80% accuracy for multi-sentence temples, 90%+ for single easy fields
- But remember: information is redundant!



# Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds  
to the correct parse of  
*“This will panic buyers !”*

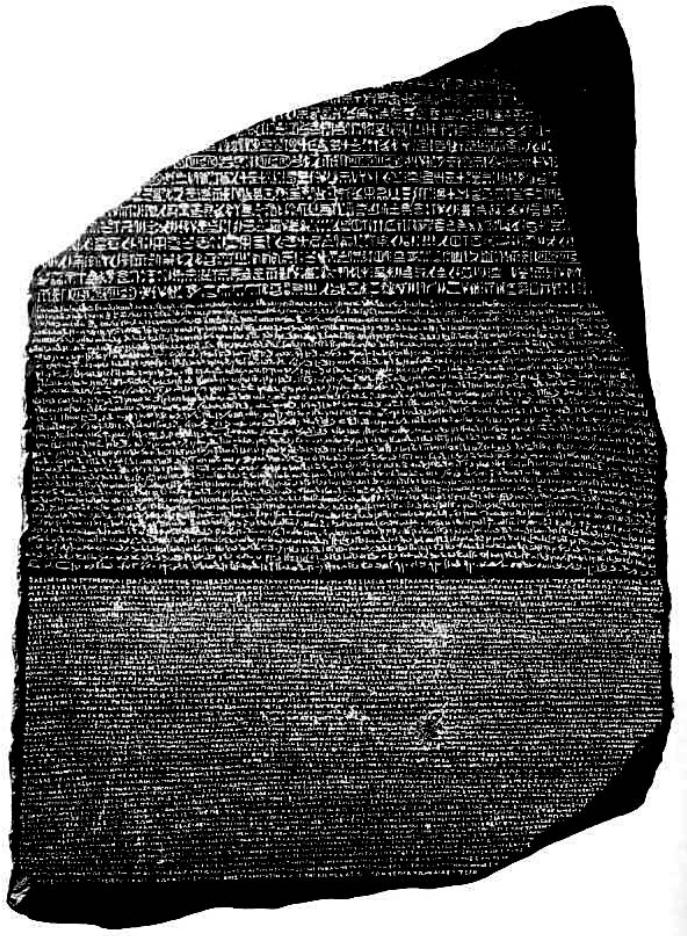


- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this



# Corpora

---



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what