



**Department of Computer Science & Engineering
(CYBER SECURITY)**

Natural Language Processing (4233OE)

Name : M.Rajeswari

Department : CSE-CS

Designation : Assist. Professor



Language Technologies



Goal: Deep Understanding

- Requires context, linguistic structure, meanings...

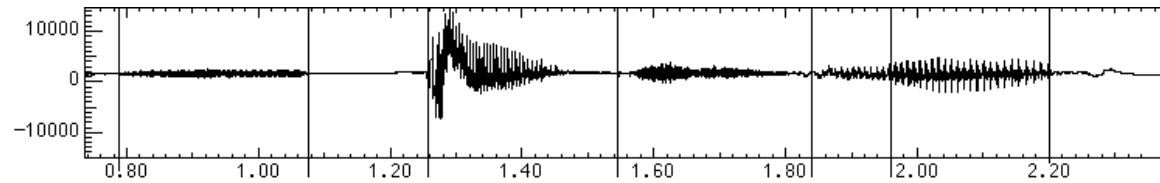
Reality: Shallow Matching

- 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

Image: Wikipedia



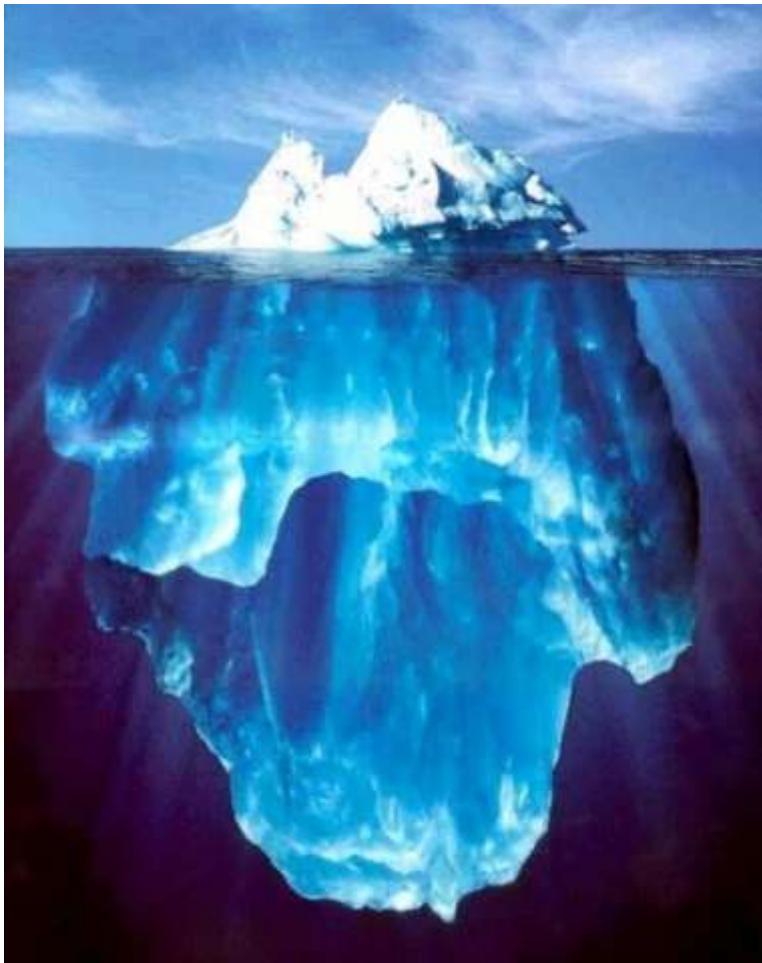
Text Data is Superficial

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.





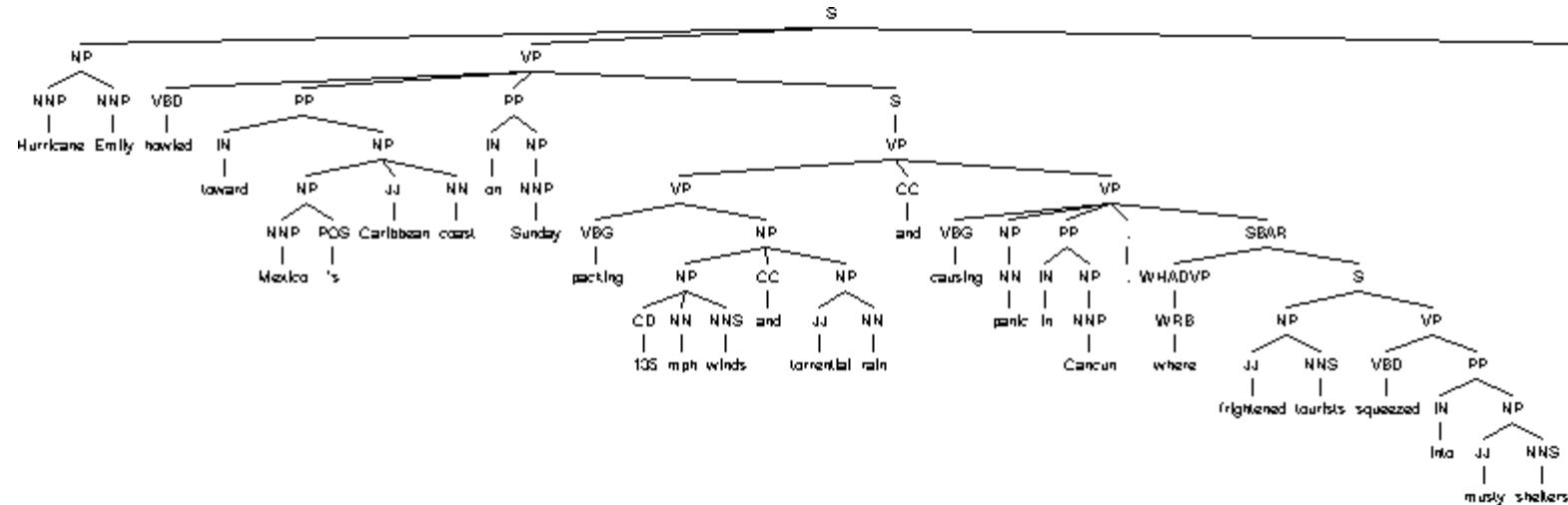
... But Language is Complex



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



Corpora

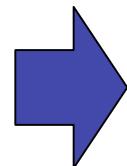
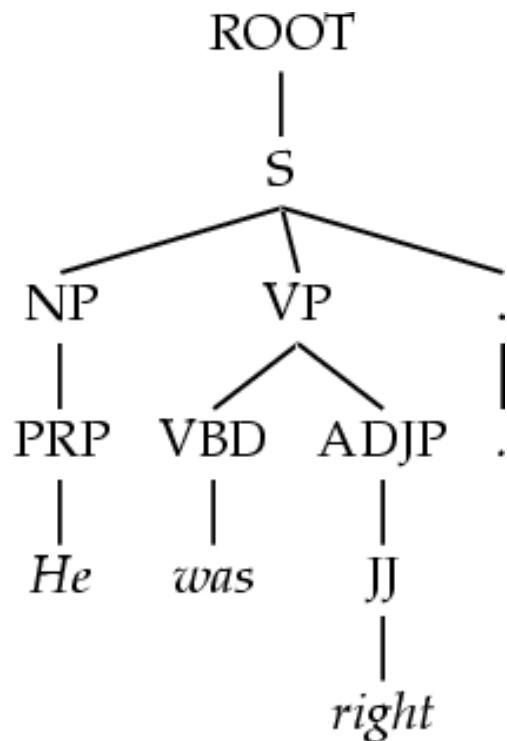


- A corpus is a collection of text
 - Often annotated in some way
 - Sometimes just lots of text
 - Balanced vs. uniform corpora
- Examples
 - Newswire collections: 500M+ words
 - Brown corpus: 1M words of tagged “balanced” text
 - Penn Treebank: 1M words of parsed WSJ
 - Canadian Hansards: 10M+ words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

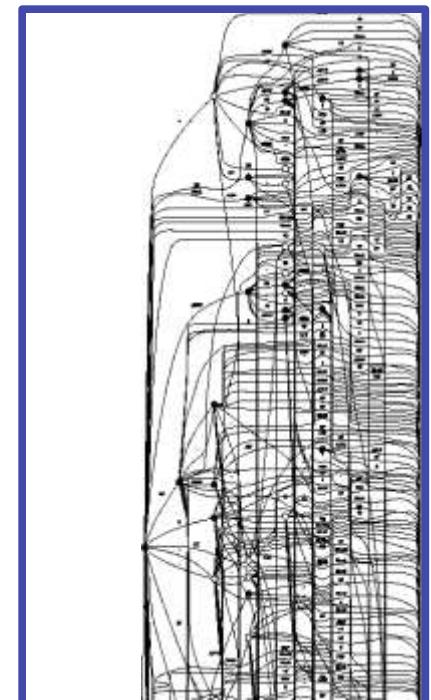
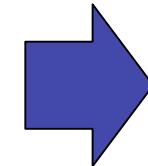


Corpus--Based Methods

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



$\text{ROOT} \rightarrow \text{S}$
 $\text{S} \rightarrow \text{NP VP} .$
 $\text{NP} \rightarrow \text{PRP}$
 $\text{VP} \rightarrow \text{VBD ADJ}$

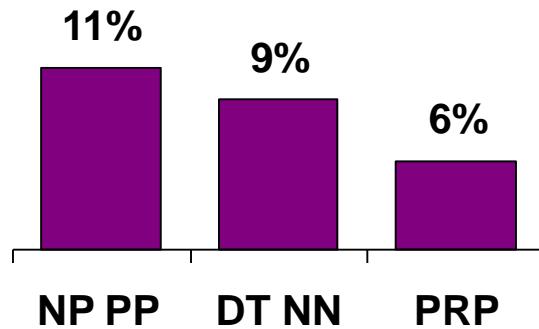




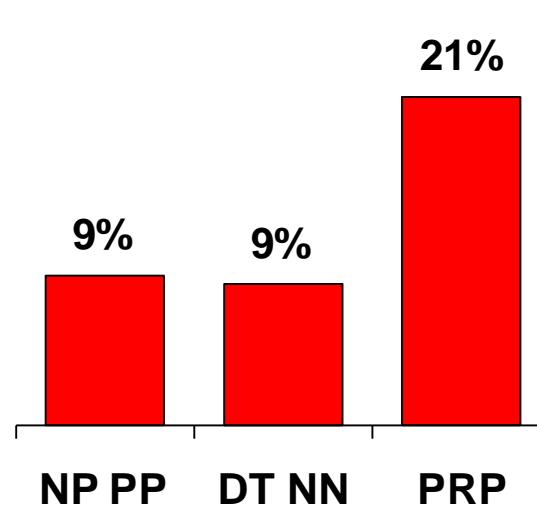
Corpus--Based Methods

- It gives us statistical information

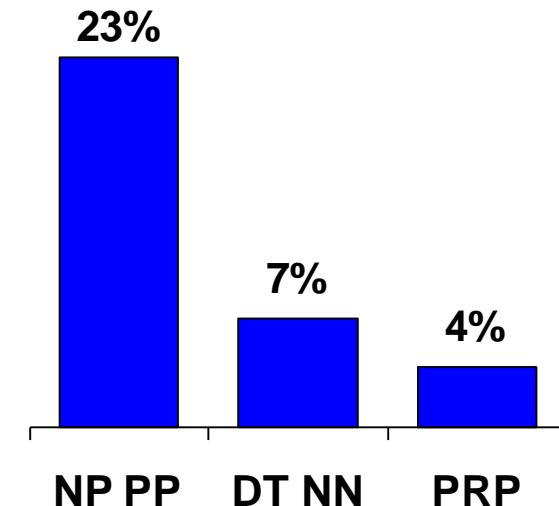
All NPs



NPs under S



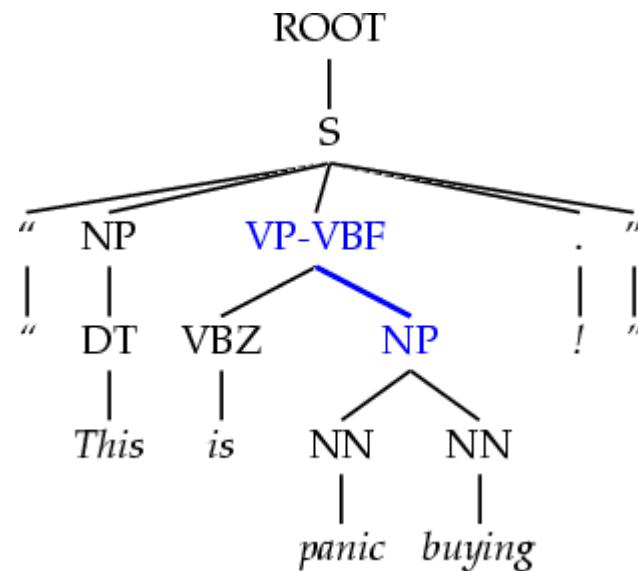
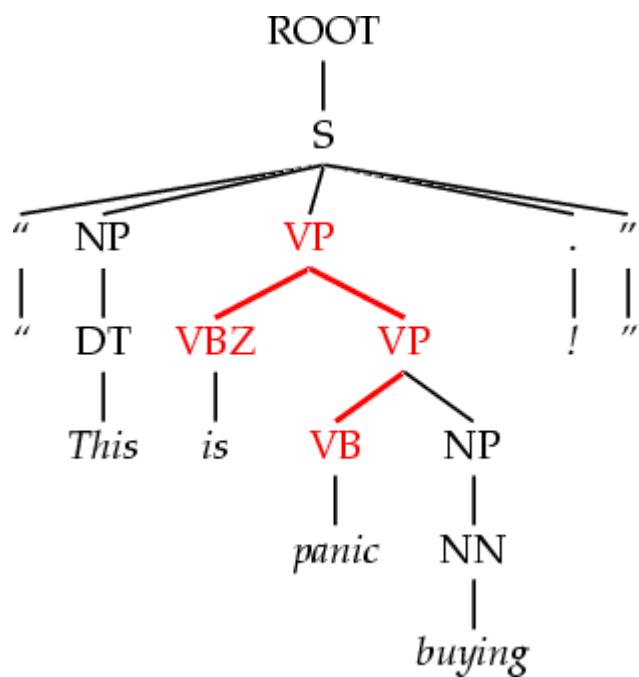
NPs under VP





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

<i>sarà</i>	<i>andata</i>
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

Google™ [Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [more »](#)

any US states' capitals are also their largest cities?

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 - 21k - [Cached](#) - [Similar pages](#)

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

Chey·enne (shē·ĕn' , -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 a multilingual tree encyclopedia

► [Des One](#) [Vogu en.w](#)

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

[Wilson P. Mendez](#)

[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

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

Entry Discussion Read Edit History Search

Log in / create account

a camel is a horse designed by a committee

Contents [hide]

1 English

1.1 Alternative forms

1.2 Proverb

The Phrase Finder

» > [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.

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

"I can assure this is not," 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 punt your butt."

Schwartz set bail at \$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 she could be held without bail.

Get a quote now. Have it, then buy online later.

Lindsay Lohan charged with theft

Join the conversation

Top Life & Style articles

Story Tools

WA Today Jobs

SUMMARY: Lindsay Lohan charged with theft

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 blonde starlet in jail. But it was the first time a judge invoked 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 dozen of assembled reporters.

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

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

Judge tells Lohan she's no star in his courtroom

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 blonde starlet in jail. But it was the first time a judge invoked 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 dozen of assembled reporters.

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

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>

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

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 de la TVA 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](#)

Google™ 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

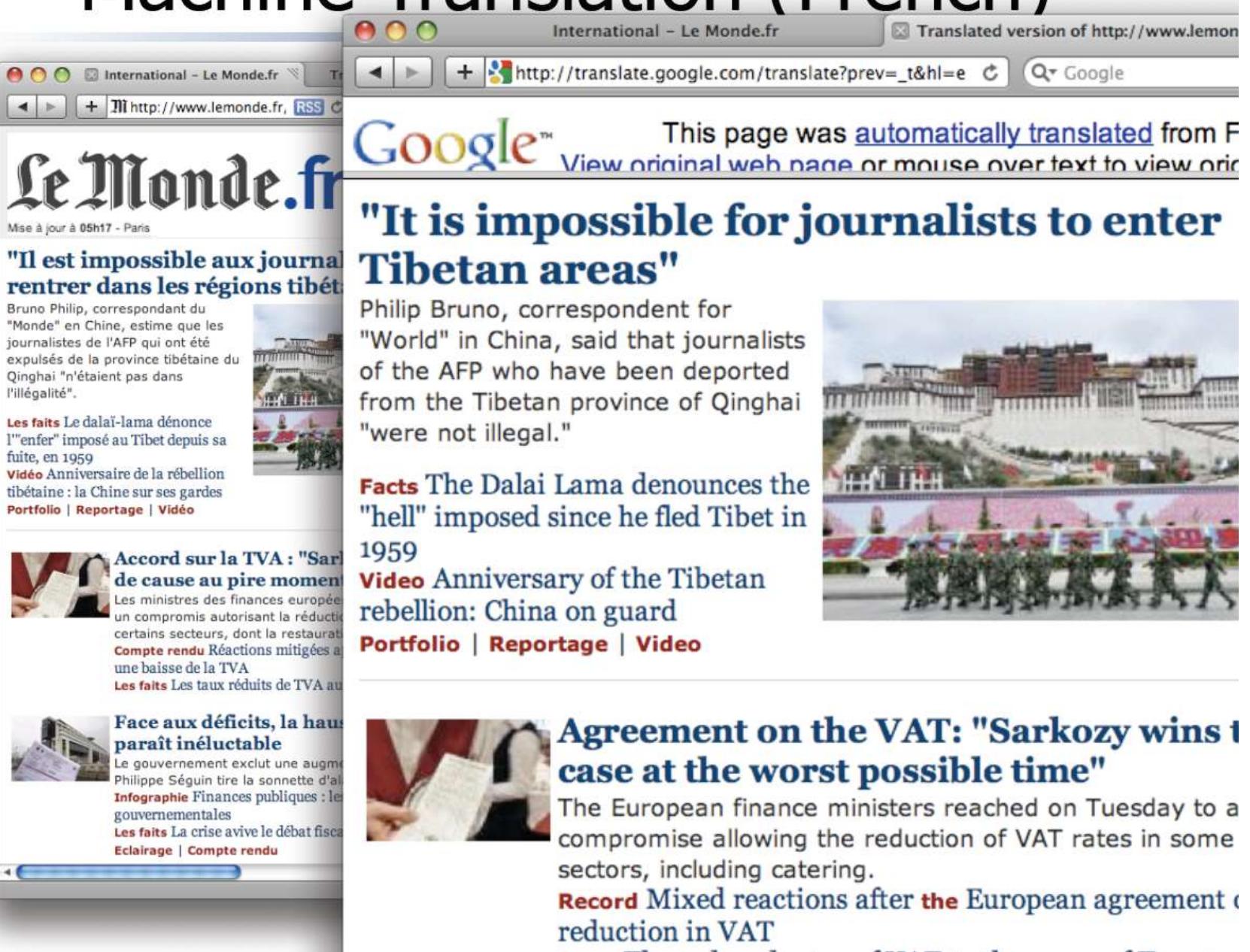
Video Anniversary of the Tibetan rebellion: China on guard

[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





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

Business

Latest News

▶ 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 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月期の国内総生産は、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

▶ 【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)



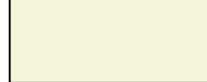
New Prius

▶ 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]



Data and Knowledge

- Classic knowledge representation worry: How will a machine ever know that...
 - Ice is frozen water?
 - Beige looks like this:

 - Chairs are solid?
- Answers:
 - 1980: write it all down
 - 2000: get by without it
 - 2020: learn it from data



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 <i>Quantum Computer Services</i>



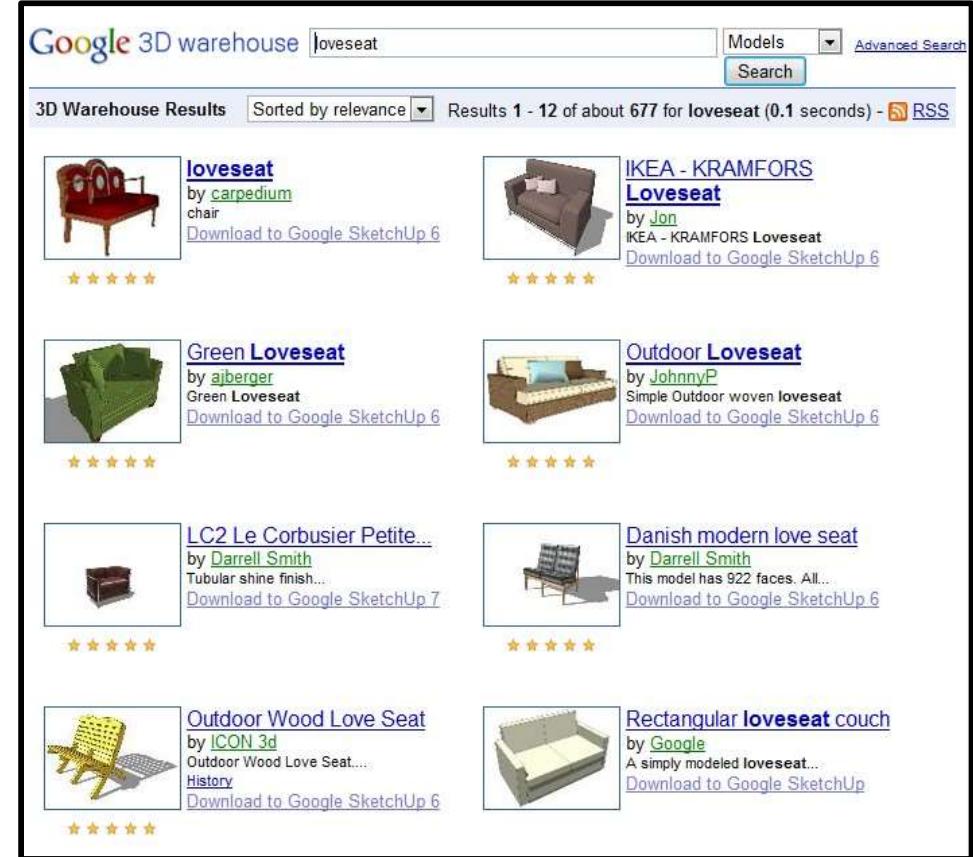
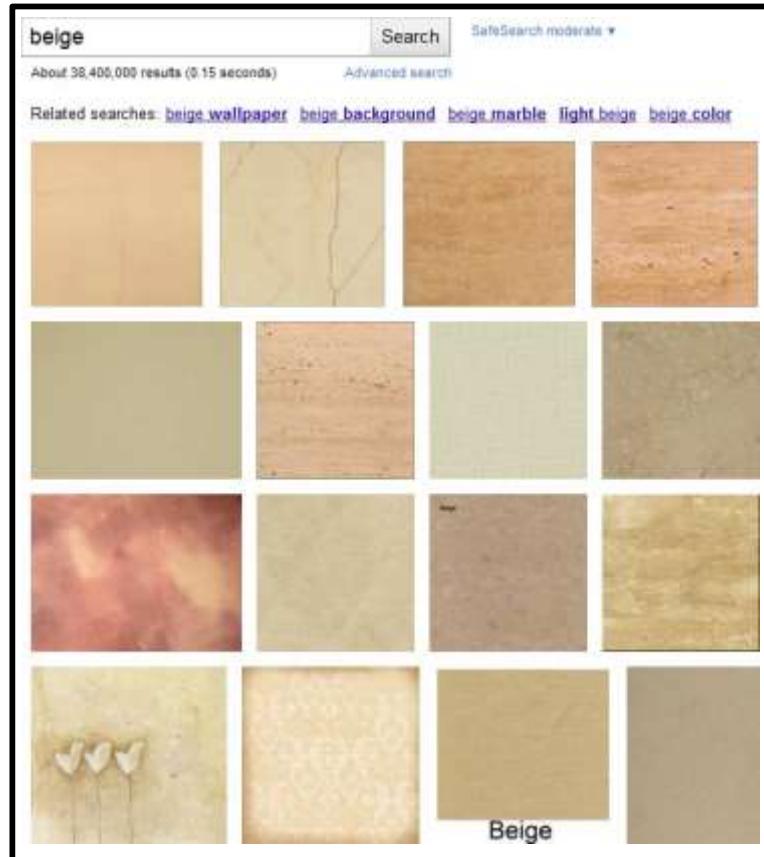
Grounded Language





Grounding with Natural Data

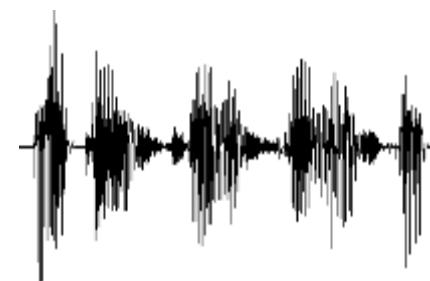
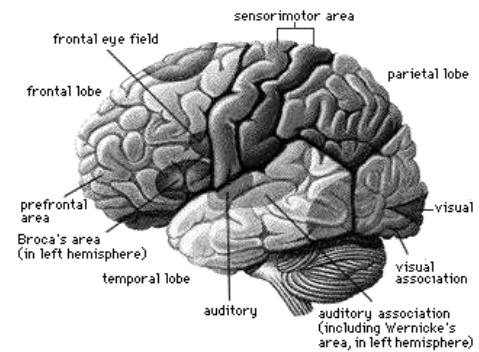
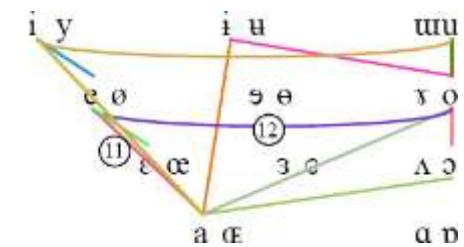
... on the beige loveseat.





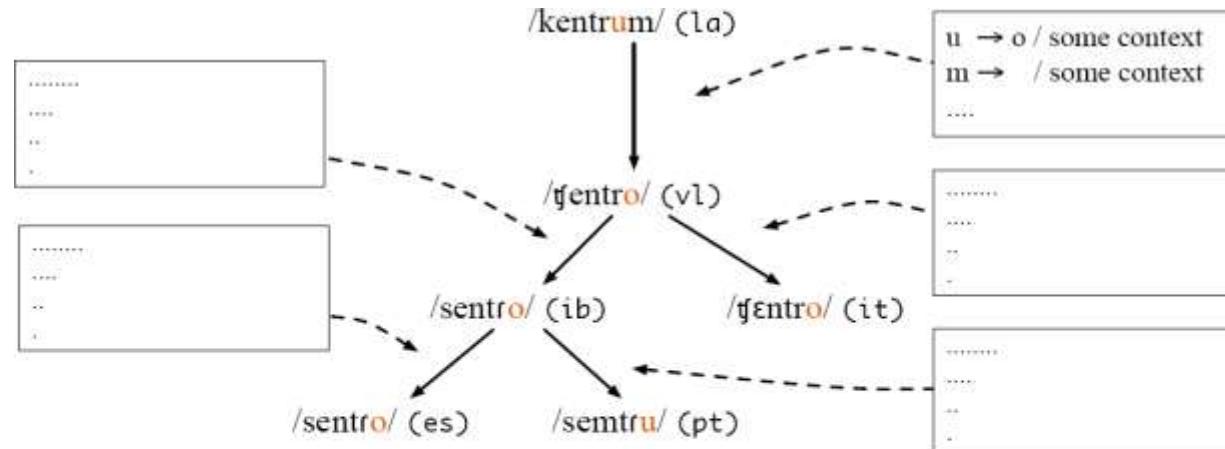
What is Nearby NLP?

- 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





Example: NLP Meets CL



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

- Example: Language change, reconstructing ancient forms, phylogenies
... just one example of the kinds of linguistic models we can build



What is NLP research?

- Three aspects we often investigate:
 - **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)



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*



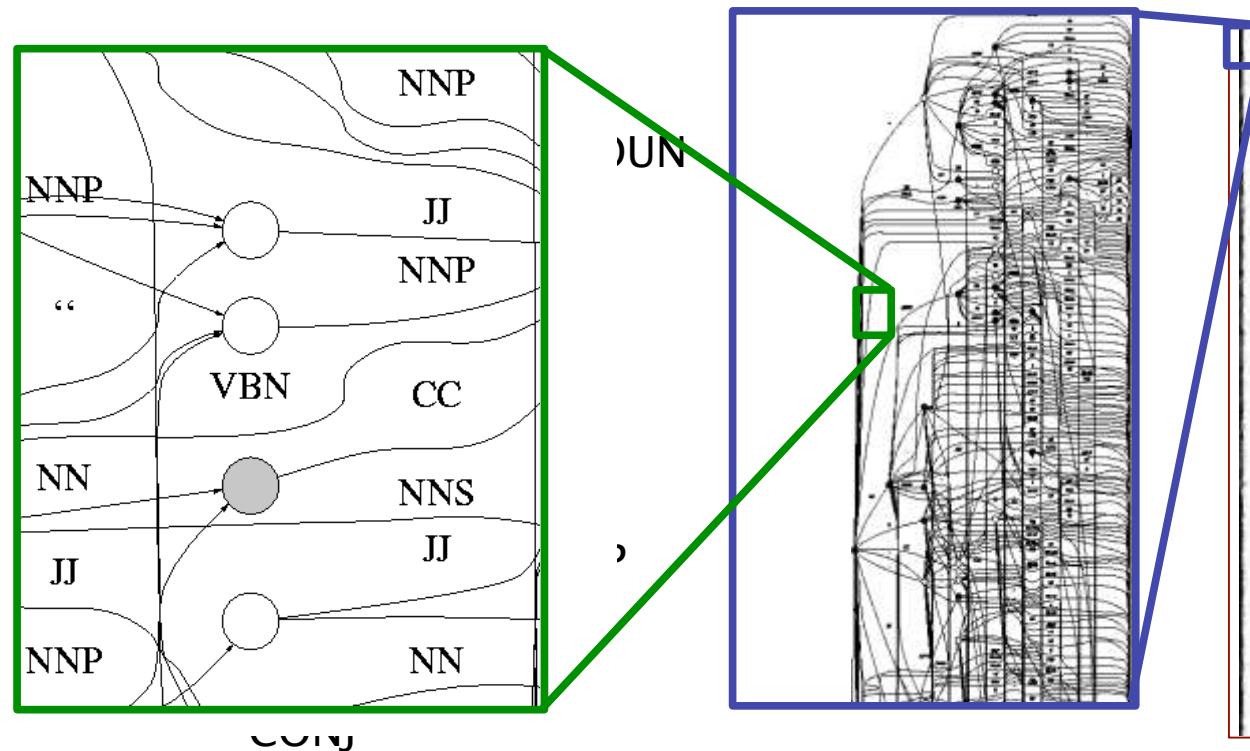
Problem: Structure

- 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)

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

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

$\text{NP} \rightarrow \text{DT NN}$

$\text{NP} \rightarrow \text{NN NNS}$

$\text{NP} \rightarrow \text{NP PP}$

$\text{VP} \rightarrow \text{VBP NP}$

$\text{VP} \rightarrow \text{VBP NP PP}$

$\text{PP} \rightarrow \text{IN NP}$

Lexicon

$\text{NN} \rightarrow \text{interest}$

$\text{NNS} \rightarrow \text{raises}$

$\text{VBP} \rightarrow \text{interest}$

$\text{VBZ} \rightarrow \text{raises}$

...

- Use deduction systems to prove parses from words

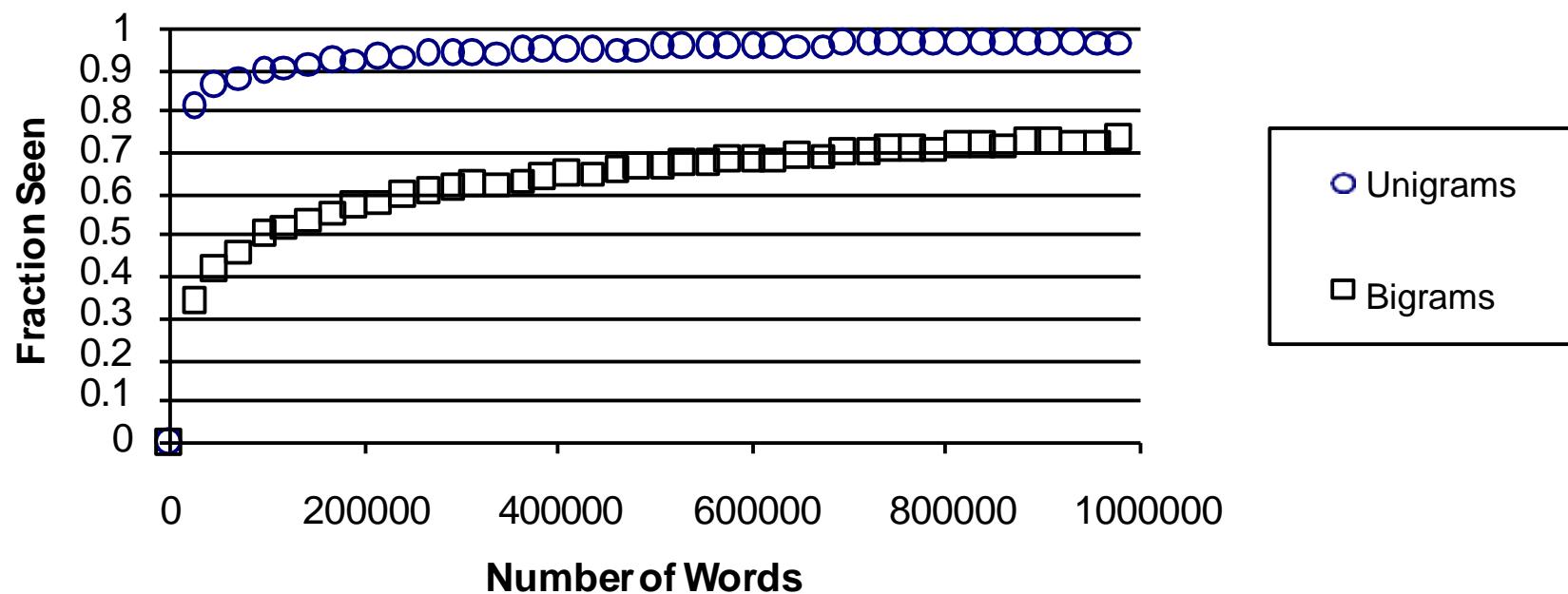
- Minimal grammar on “Fed raises” sentence: 36 parses
- Simple 10-rule grammar: 592 parses
- Real-size grammar: many millions of parses

- 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





The (Effective) NLP Cycle

- Pick a problem (usually some disambiguation)
- Get a lot of data (usually a labeled corpus)
- Build the simplest thing that could possibly work
- Repeat:
 - Examine the most common errors are
 - Figure out what information a human might use to avoid them
 - Modify the system to exploit that information
 - Feature engineering
 - Representation redesign
 - Different machine learning methods
- We're do this over and over again