

Course Overview

Lecture #1

Computational Linguistics

CMPSCI 585, Fall 2007

University of Massachusetts Amherst



Andrew McCallum

`http://www.cs.umass.edu/~mccallum/courses/inlp2007`

Where to find syllabus, announcements,
slides, homeworks

Today's Main Points

- Why is natural language interesting and difficult, complex and ambiguous.
 - Why? How to we resolve this ambiguity?
- Six “layers” of natural language
- Natural Language Processing overview, current successes
- Get to know each other, and our motivations for being here
- Course mechanics; what you can expect

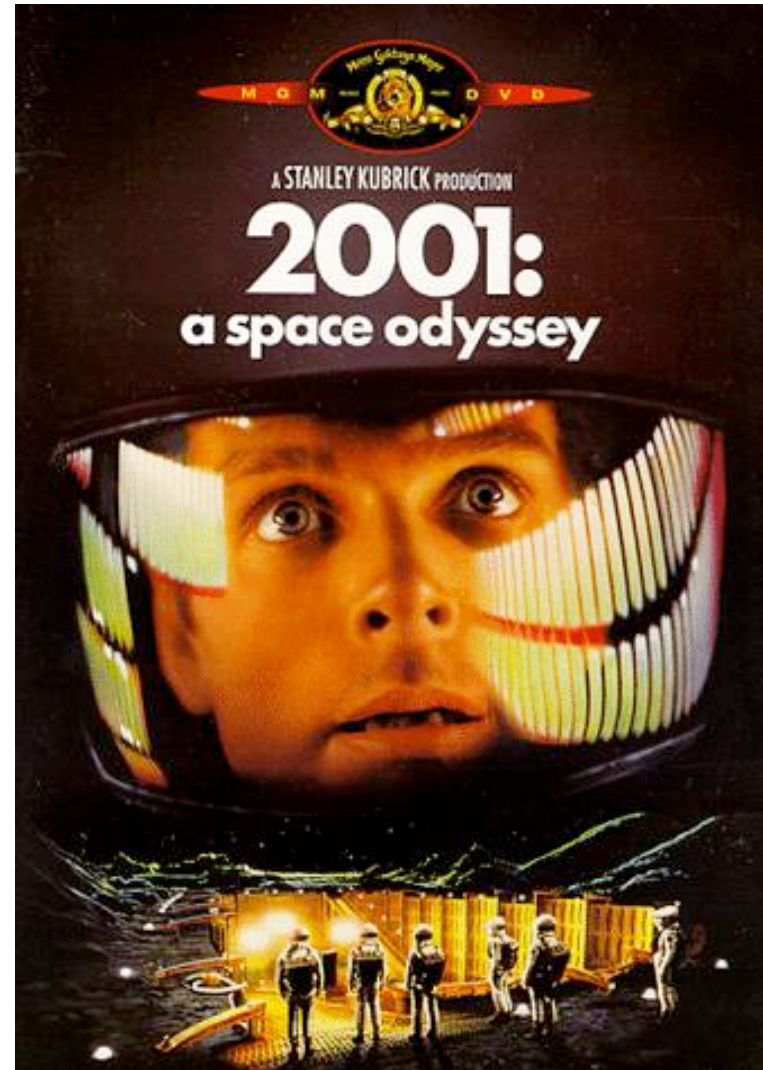
1967



*Stanley Kubrick,
filmmaker
1928 - 1999*

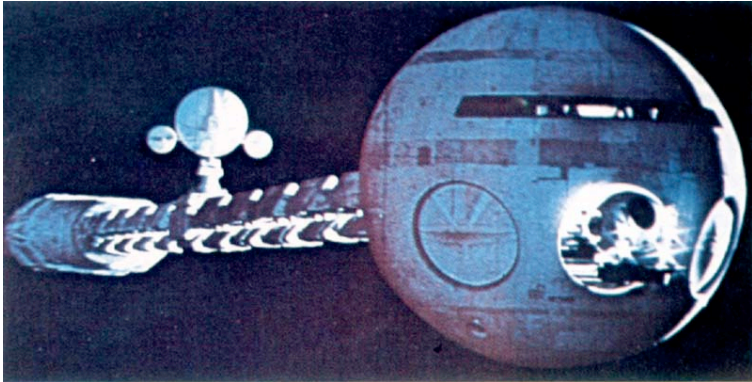


*Arthur C. Clarke,
author, futurist,
1917 -*



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

HAL



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

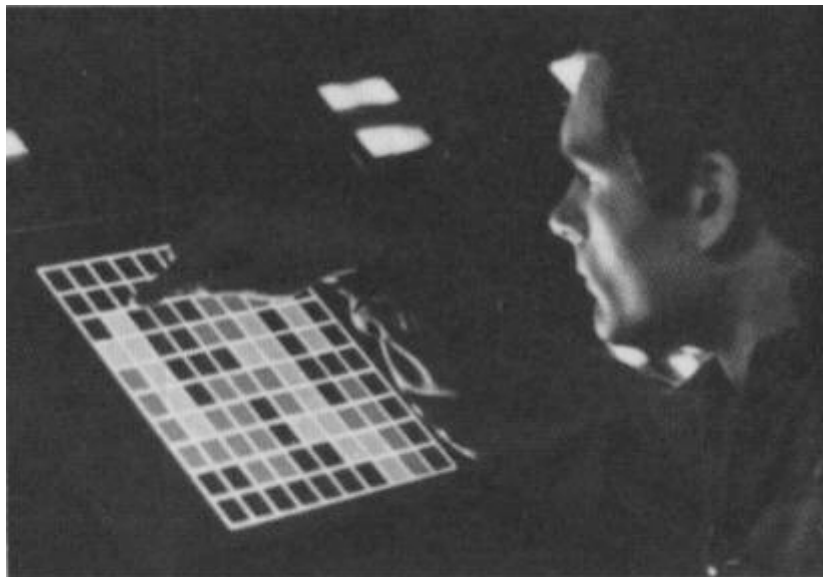
HAL's Capabilities

- Display graphics
- Play chess
- *Natural language production and understanding*

- Vision
- Planning
- Learning
- ...

Graphics

HAL

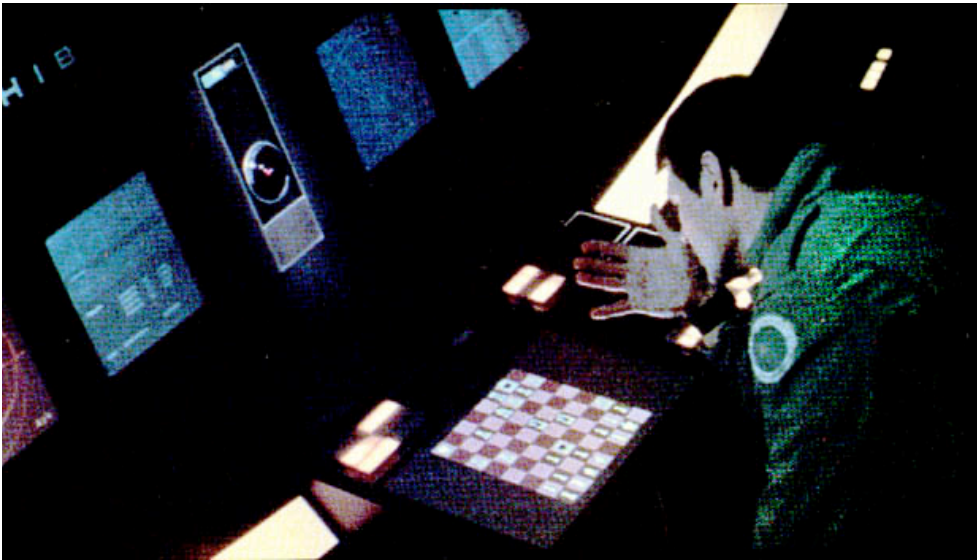


Now

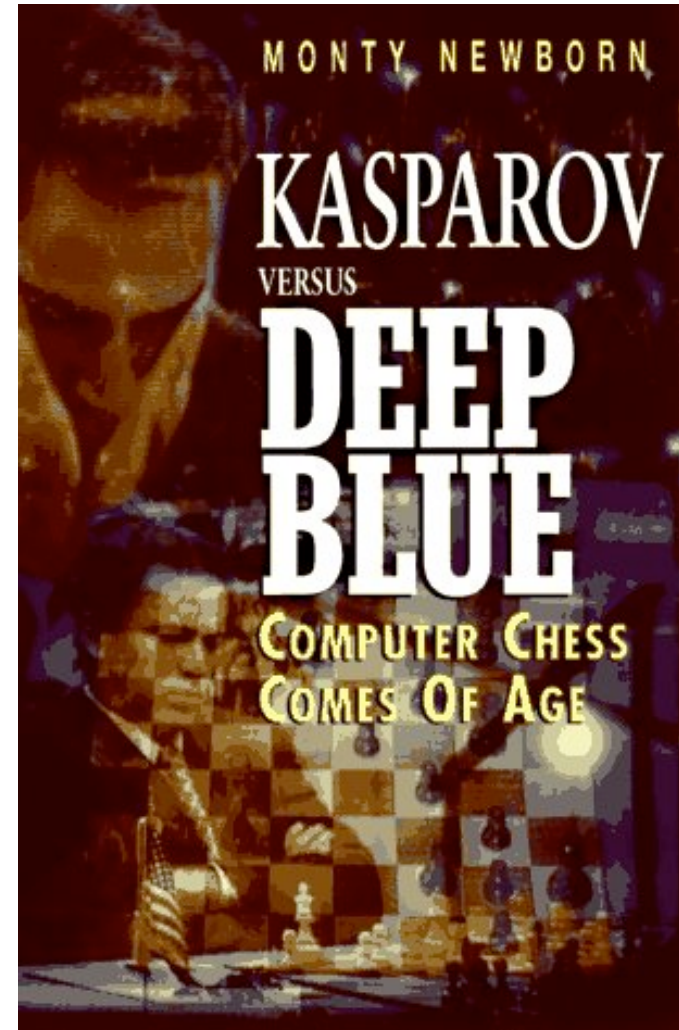


Chess

HAL



Now



Natural Language Understanding

HAL

David Bowman:

Open the pod bay doors, Hal.

HAL:

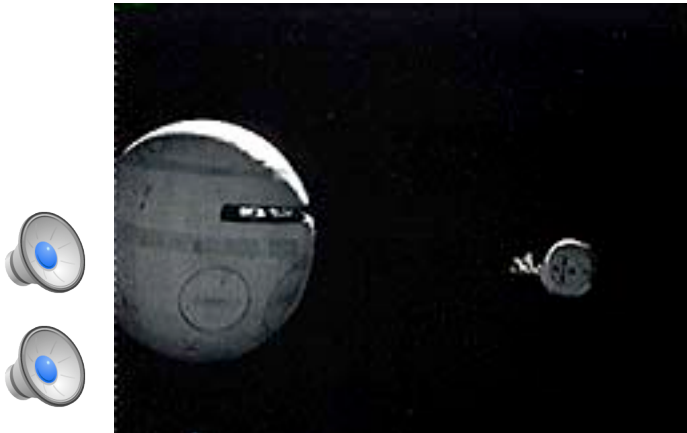
I'm sorry, Dave, I'm afraid I can't do that.

David Bowman:

What are you talking about, Hal?

...HAL:

I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



Andrew McCallum, UMass Amherst,
including material from Chris Manning and Jason Eisner

Now



Many useful tools, but none that come even close to HAL's ability to communicate in natural language.

1950



Alan Turing
1912 - 1954

Turing Test

“Computing Machinery and Intelligence”
Mind, Vol. 59, No. 236, pp. 433-460, 1950

I propose to consider the question
"Can machines think?" ...
We can only see a short distance ahead, but
we can see plenty there that needs to be done.

Layers of Computational Linguistics

1. Phonetics & Phonology
2. Morphology
3. Syntax
4. Semantics
5. Pragmatics
6. Discourse

1. Phonetics & Phonology

The study of: language sounds,
how they are
physically formed;

systems of discrete
sounds, e.g. languages'
syllable structure.

dis-k&- 'nekt

disconnect

“It is easy to recognize speech.”

“It is easy to wreck a nice beach.”

JeetJet?

2. Morphology

The study of the sub-word units of meaning.

disconnect

“not”

“to attach”

Even more necessary in some other languages,
e.g. Turkish:

uygarlastiramadiklarimizdanmissinizcasina

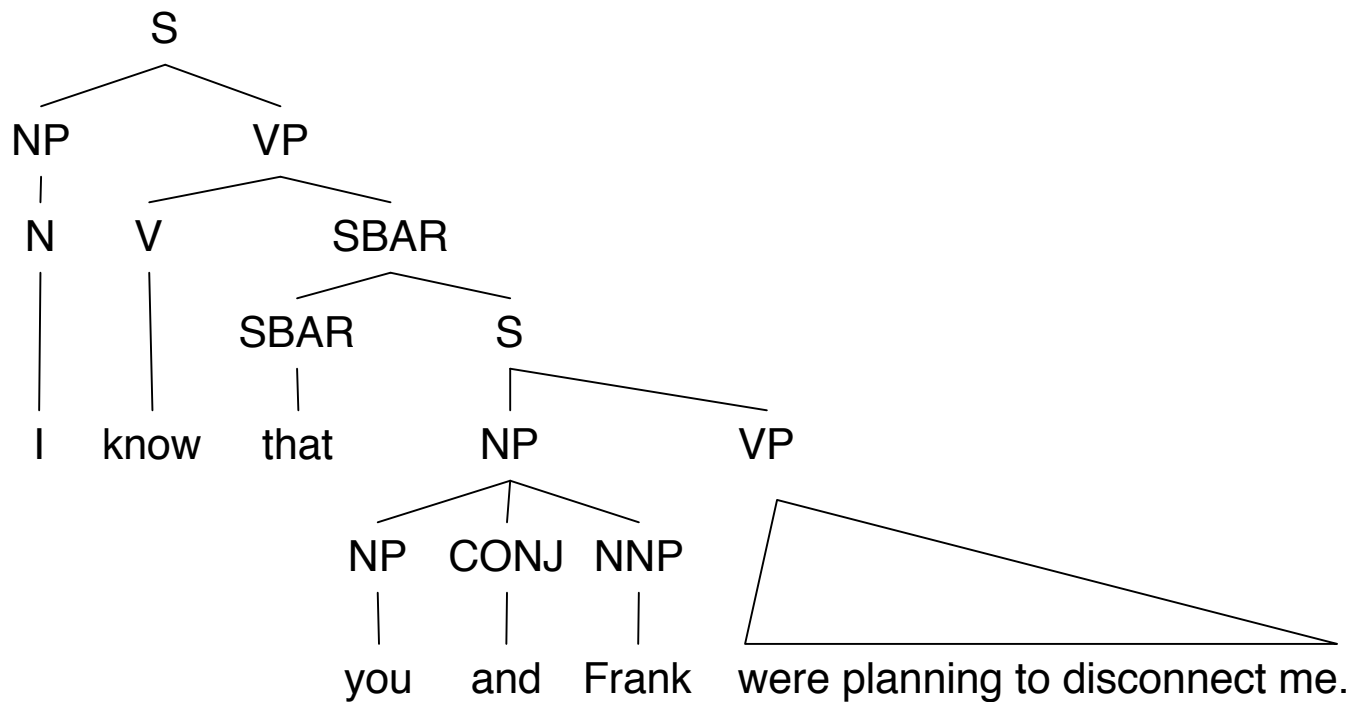
uygar las tir ama dik lar imiz dan mis siniz casina

(behaving) as if you are among those whom we could not civilize

3. Syntax

The study of the structural relationships between words.

I know that you and Frank were planning to disconnect me.



Not same structure:

You know me--Frank and I were planning to disconnect that.

4. Semantics

The study of the literal meaning.

I know that you and Frank were planning to disconnect me.

ACTION = disconnect

ACTOR = you and Frank

OBJECT = me

5. Pragmatics

The study of how language is used to accomplish goals.

What should you conclude from the fact I said something?
How should you react?

I'm sorry Dave, I'm afraid I can't do that.

Includes notions of polite and indirect styles.

6. Discourse

The study of linguistic units larger than a single utterance.

The structure of conversations:
turn taking, thread of meaning.

David Bowman:

Open the pod bay doors, Hal.

HAL:

I'm sorry, Dave, I'm afraid I can't do that.

David Bowman:

What are you talking about, Hal?

...HAL:

I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

Linguistic Rules

E.g. Morphology

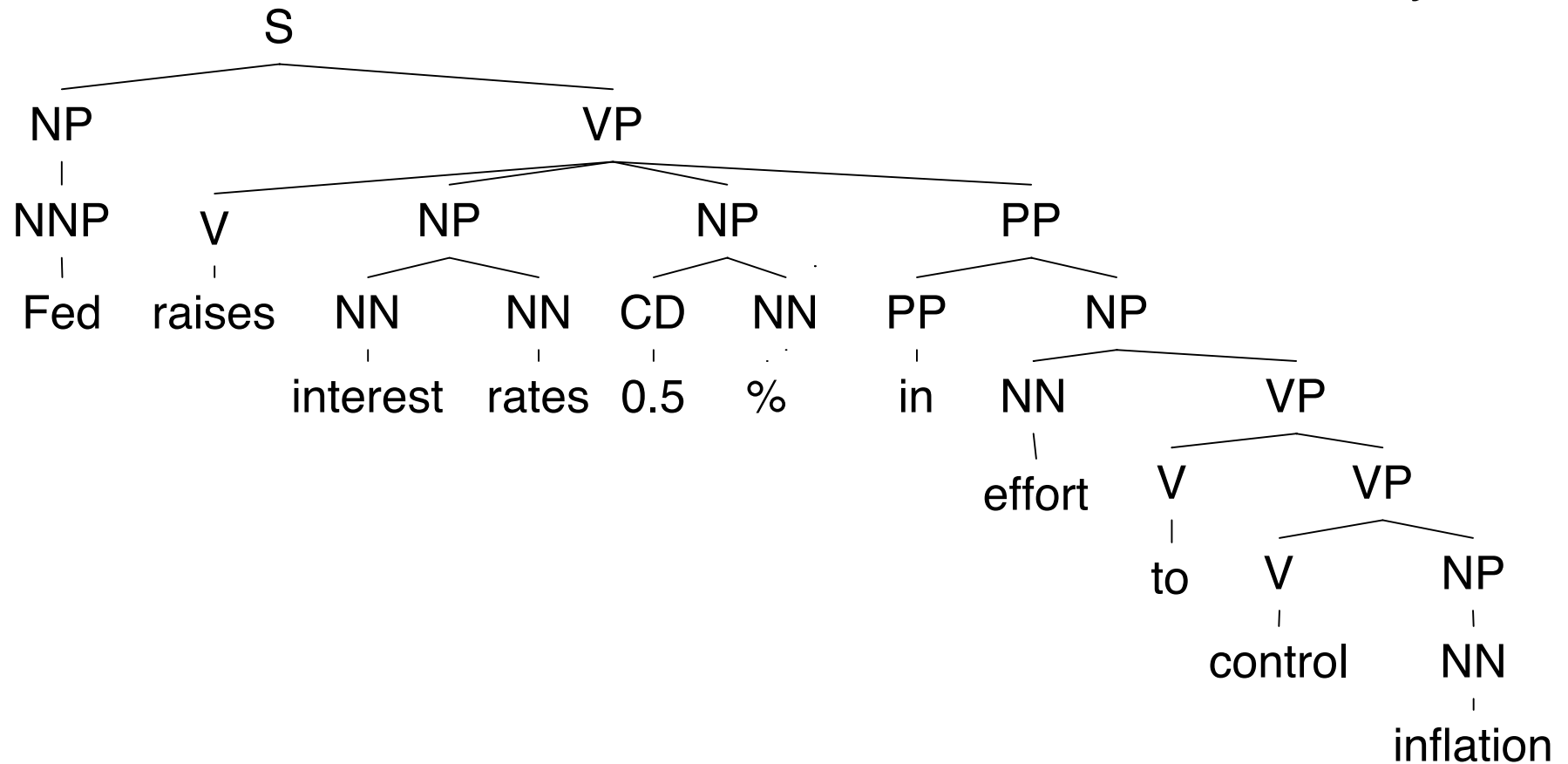
To make a word plural, add “s”

- dog → dogs
- baby → babies
- dish → dishes
- goose → geese
- child → children
- fish → fish (!)

Inherent Ambiguity in Syntax

Fed raises interest rates 0.5%
in effort to control inflation

NY Times headline 17 May 2000



Where are the ambiguities?

Part-of-speech ambiguities

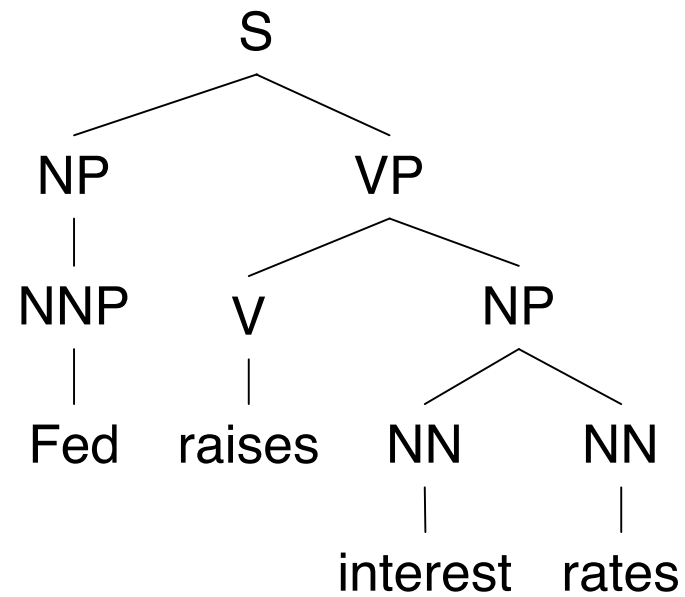
Syntactic attachment ambiguities

		VB				
	VBZ	VBZ	VBZ			
NNP	NNS	NNS	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

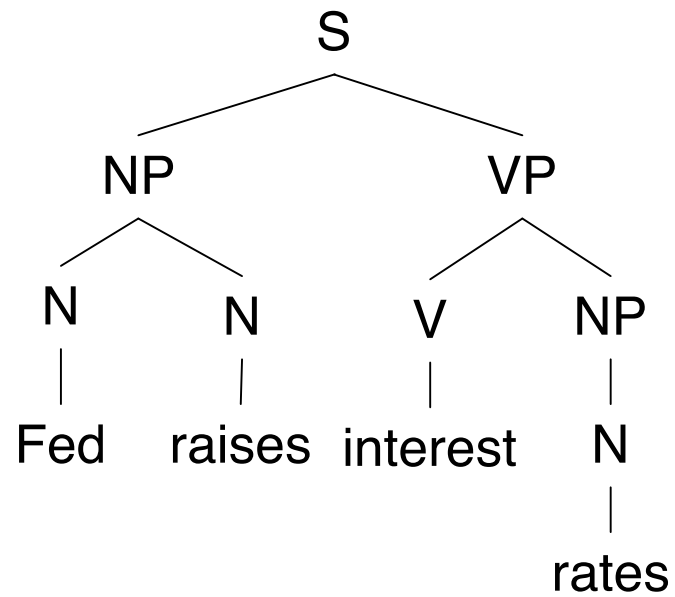
Word sense ambiguities: Fed → "federal agent"
interest → a feeling of wanting to know or learn more

Semantic interpretation ambiguities above the word level.

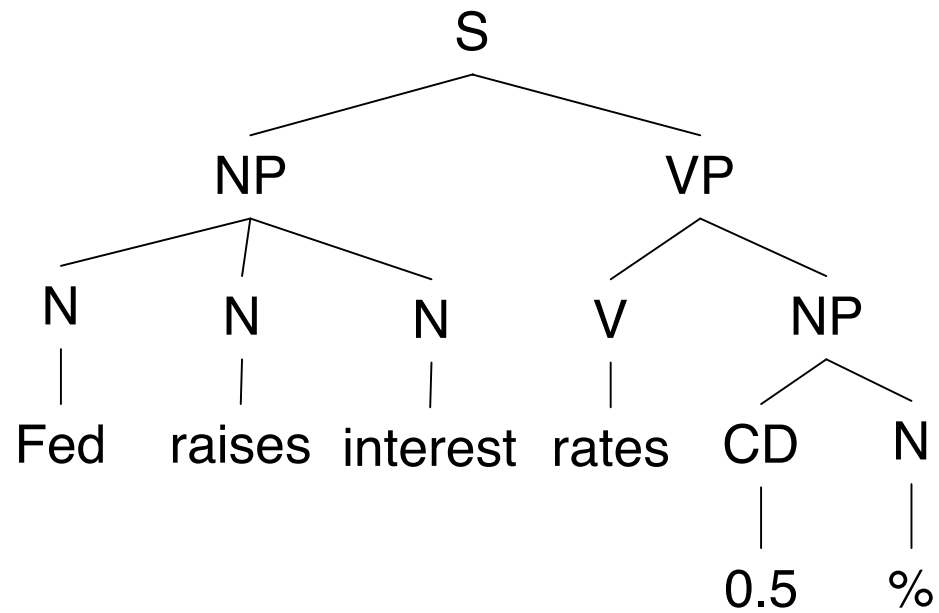
Effects of V/N Ambiguity (1)



Effects of V/N Ambiguity (2)



Effects of V/N Ambiguity (3)



Ambiguous Headlines

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Ban on Nude Dancing on Governor's Desk

Language Evolves

- Morphology
 - We learn new words all the time:
bioterrorism, cyberstalker, infotainment,
thumb candy, energy bar
- Part-of-speech
 - Historically: “kind” and “sort” were always *nouns*:
“I knowe that sorte of men ryght well.” [1560]
 - Now also used as *degree modifiers*:
“I’m sort of hungry.” [Present]
“It sort o’ stirs one up to hear about old times.” [1833]

Natural Language Computing is hard because

- Natural language is:
 - highly ambiguous at all levels
 - complex and subtle
 - fuzzy, probabilistic
 - interpretation involves ***combining evidence***
 - involves reasoning about the world
 - embedded a social system of people interacting
 - persuading, insulting and amusing them
 - changing over time

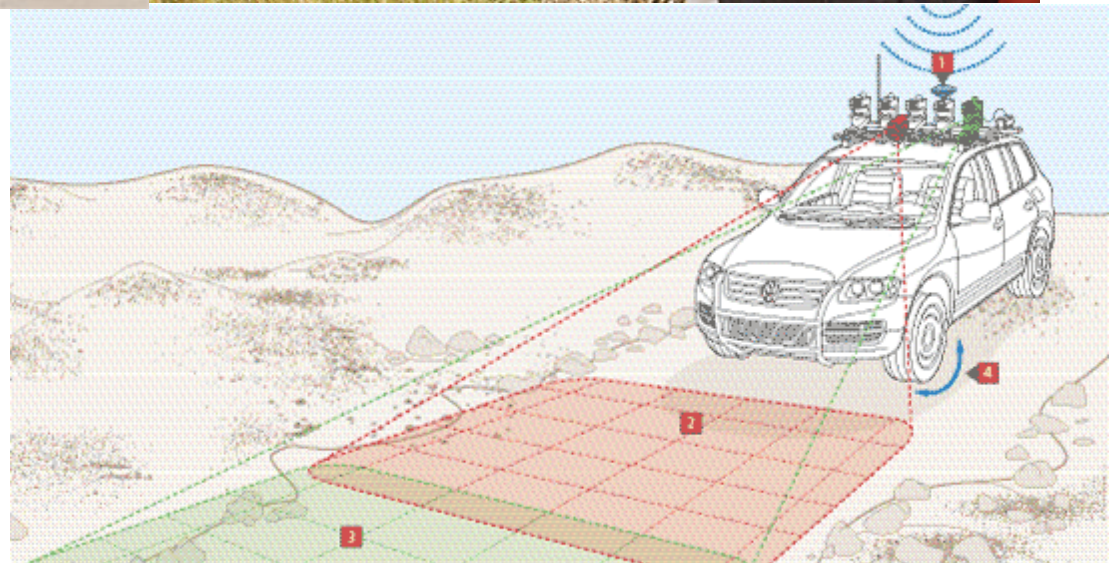
Probabilistic Models of Language

To handle this **ambiguity** and to **integrate evidence** from multiple levels we turn to:

The tools of probability:

- Bayesian Classifiers (not rules)
- Hidden Markov Models (not DFAs)
- *Probabilistic* Context Free Grammars
- ...other tools of Machine Learning, AI, Statistics

Another Area where Probabilistic Combination of Evidence Won



Andrew McCallum, UMass Amherst, including material from Chris Manning and Jason Eisner

Natural Language Processing

- Natural Language Processing (NLP) studies how to get computers to do useful things with natural languages:
 - Most commonly Natural Language Understanding
 - The complementary task is Natural Language Generation
- NLP draws on research in Linguistics, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.

Engineering
Goal

What & Where is NLP

- Goals can be very far-reaching
 - True text understanding
 - Reasoning and decision-making from text
 - Real-time spoken dialog
- Or very down-to-earth
 - Searching the Web
 - Context-sensitive spelling correction
 - Analyzing reading-level or authorship statistically
 - Extracting company names and locations from news articles.
- These days, the later predominate (as NLP becomes increasingly practical, focused on performing measurably useful tasks *now*).
- Although language is complex, and ambiguity is pervasive, NLP can also be surprisingly easy sometimes:
 - rough text features often do half the job

Linguistics

- Linguistics is the study of natural languages:
 - Understanding this naturally-occurring phenomenon.
 - Structure, meaning, how acquired, differences and commonalities across languages.
- Linguistics draws on research in Natural Language Processing, Theoretical Computer Science, Artificial Intelligence, Mathematics and Statistics, Psychology, Cognitive Science, etc.

Scientific
Goal

Example Applications of NLP

The screenshot shows a web browser window with the Google search engine. The search query is "natural language processing". The results page displays several search results, including a sponsored link for "Natural Language Search" and a link for "Work at Google".

Google Search: natural language processing

http://www.google.com/search?hl=en&ie=ISO-8859-1&ec... Google

Yahoo! Google Slashdot News McC Research Reviewing Mac Java Thesaurus Funding Teaching

Google Search: natural I... http://www.canberra.ed...

Google™ natural language processing Google Search

Advanced Search Preferences Language Tools Search Tips

Web Images Groups Directory News

Searched the web for natural language processing. Results 1 - 10 of about 2,210,000. Search took 0.21 seconds.

Natural Language Processing
Natural Language Processing should make it possible for people to use computers in much the same way that they would use a human assistant to get their work ...
research.microsoft.com/nlp/ - 28k - [Cached](#) - [Similar pages](#)

ISI's Natural Language Group
Overview of Research Environment **Natural Language Processing** at USC/ISI. ... USC offers a wide range of courses in areas related to **natural language processing**. ...
Description: The **Natural Language Processing** group at the Information Sciences Institute of the University of Southern...
Category: [Computers](#) > [Artificial Intelligence](#) > ... > [Research Groups](#)
www.isi.edu/natural-language/nlp-at-isi.html - 15k - [Cached](#) - [Similar pages](#)

Foundations of Statistical Natural Language Processing
Foundations of Statistical **Natural Language Processing**. ... Chris Manning and Hinrich Schütze, Foundations of Statistical **Natural Language Processing**, MIT Press. ...
nlp.stanford.edu/fsnlp/ - 7k - [Cached](#) - [Similar pages](#)

Yahoo! Directory Artificial Intelligence > Natural Language ...
Artificial Intelligence > **Natural Language Processing** Directory > Science > Computer Science > Artificial Intelligence > **Natural Language Processing**, ...
dir.yahoo.com/Science/Computer_Science/Artificial_Intelligence/Natural_Language_Processing/

Sponsored Links

Natural Language Search
Returns more relevant searches
Installs in days. Free White Papers
www.primus.com
Interest:

NLP News
All the news that's fit to parse
Human Language Technology
fieldmethods.net
Interest:

Natural Lang. Processing:
Text Mining Tool Based on NLP -
For scientific literature analysis.
www.ariadnegenomics.com
Interest:

Work at Google
Google is hiring expert computer
scientists and software developers

Example Applications of NLP: MSWord spelling correction, grammar checking

If you use Microsoft Word you have no doubt noticed red any misspelled words (or, to be exact, all words that did you know that you can correct these errors simply Microsoft Word will give you a list of the words that if word you want appaers in the list) you simply pick it fi



Example Applications of NLP:

The screenshot shows the Google News homepage in a Firefox browser window. The address bar displays the URL: <http://news.google.com/?auth=DQAAAG4AAAAtv1ZP23e3kAneETvc5X1kNI-5CN0uomVikJWB>. The page features the Google News logo, navigation links for Web, Images, Groups, News, Froogle, Local, and more, along with search boxes for News and the Web. A yellow banner highlights that search history now includes Google News. The main content area is divided into sections: Top Stories (U.S.), Personalize this page, and In The News. The Top Stories section includes articles such as "Alito Seen as Carrying the Torch of Reagan" and "Quotes: Remembering Coretta Scott King". The Personalize this page section lists related news items like "Rates to rise as Fed balances act" and "Madonna Meets Skype". The In The News section features links to news about Bob Woodruff, Coretta Scott King, Wendy Wasserstein, Exxon Mobil, Doug Vogt, Jill Carroll, Buick Invitational, State of the Union, Saddam Hussein, and Screen Actors Guild. The bottom of the page shows a "World" section with an article "India likely to vote against Iran" and a "U.S." section with an article "Injured ABC newsmen return to US".

Example Applications of NLP

Information Extraction: Find experts, employees

http://www.zoominfo.com

Dr. Andrew McCallum
Action Editor
Journal of Machine Learning Research
Last Mentioned on 10/12/2003

Actions

- [Send This Profile](#)
- [Update Your Profile](#)
- [Email Not Available](#)

Other Titles Held:

Member, Editorial Board

Additional Current Employment

Carnegie Mellon University	Post-Doctoral Fellow
	Adjunct Faculty Member
	Adjunct Faculty Position
University of Massachusetts Amherst, CO	Research Associate Professor
Adjunct Faculty	Research Scientist

Board Memberships and Affiliations

Intelliseek Inc	Member of Advisory Board
IJCAI	Member, Program Committees (past)
AAAI	Member, Program Committees (past)
ICML	Member, Program Committees (past)
NIPS	Member, Program Committees (past)

Past Employment History

WhizBang Labs Inc	Vice President of Research and Development
Just Research	Research Scientist
Biomedical Information Communication Center of Oregon Health Sciences University	Machine Learning Researcher

Education

University of Rochester	Ph.D.	Computer Science
Dartmouth College	Bachelor of Arts	Computer Science

Information about Andrew McCallum was compiled from 6 sources:

[JMLR Inc](#)
<http://www.jmlr.org>

JMLR, which publishes high-quality scholarly articles in all areas of machine learning, competes with the commercial journal Machine Learning, which costs US\$1006. A number of Machine Learning editorial board members have resigned to join the editorial board of JMLR. ... [\(more\)](#)

[Click here](#) to find other people who work for *JMLR Inc*

[WhizBang Labs Inc](#)
Contact Us Corporate Headquarters
3210 North Canyon Road Suite 200
Provo, UT 84604
Phone: (801) 418-7100
Fax: (801) 818-0300

<http://www.whizbanglabs.com>

WhizBang! Labs, founded in 1999, is a leader in the field of information extraction and document auto-tagging from unstructured data sources.
Through our products and services, we analyze unstructured content in both on-line and off-line formats, locate and extract key data elements into XML-tagged ... [\(more\)](#)

[Click here](#) to find other people who work for *WhizBang Labs Inc*

[Intelliseek Inc](#)
1128 Main Street , Fourth Floor
Cincinnati, OH 45202-7236
Phone: 513-618-6700

Go to "http://networking2.eliyon.com/Networking/default.asp"

Example Applications of NLP: Information Extraction: Job Openings

foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1

Ice Cream Guru

If you dream of cold creamy chocolate or coochy boochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship. Contact Susana e-mail 1-800-488-2611

Example Applications of NLP: Information Extraction: Job Openings

The screenshot shows the FlipDog.com website interface. The browser title is "job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer". The address bar shows "http://www.flipdog.com/home.html". The website features a navigation menu with "Home", "Find Jobs", "Your Account", "Resource Center", "Support", and "Employers". A banner at the top reads "Job Search at FlipDog.com: Employment & Career Management".

The main content area includes a large image of a dog with green spots, next to the text "647,514 Job Opportunities from 53,641 Employers". Below this are buttons for "Find a Job!" and "Post Your Resume". There is also a section for "Employers" with a "click here for Products & Services" link.

On the right side, there are several sections:

- Pigskin Places:** A list of job categories with counts: Health Care in NY (2,770), Health Care in MD (1,262), Sales in NY (3,751), Sales in MD (958), Computing in NY (8,050), and Computing in MD (4,114).
- Jobs for Sports Fans:** A list of job titles: Head Football Coach, Football Coach, Asst. Football Coach, High School Football Coach, and Univ. Asst. Football Coach.
- Showcase Jobs:** A section for MRI Management Recruiters of Charlotte North, featuring a logo and a description of their services. A "Learn More" link is provided.
- Job Seeker Newsletter:** A form to "Enter your e-mail address:" with a "Sign Me Up!" button.

At the bottom left, there is a "Job Seekers: Find your dream job!" section with a list of links for reports, accounts, job hunting, employer database, expert advice, salary surveys, and desktop search tools.

At the bottom right, there are several award logos: "Top 100 Web Sites" from PC Magazine (Nov. 2000), "Top 10 Career Web Site" from Media Metrix (Sept. 2000), and "Top 10 Job Site". A "powered by WhizBang!" logo is also present.

The taskbar at the bottom shows the Start button, several open applications (Microsoft PowerPoint, job search find employmen...), and the system clock showing 12:12 AM.

Example Applications of NLP: Automatically Solving Crossword Puzzles

The screenshot shows the OneAcross website interface. At the top, there is a crossword puzzle grid with the word "ONE" in the first row and "ACROSS" in the second row. To the right of the grid, there is a small grid with a blue square containing an exclamation mark and a white square containing an at-sign and a question mark. Below the grid is a navigation menu with links: Home, Crosswords, Cryptograms, Anagrams, Reference, Forum, Languages, News!, Contact@, Support, Advertise, Privacy, About. There is an Amazon.com advertisement for "new and used DVDs". A section titled "OneAcross Changes" explains that the site is moving to a better server. A "Support this Site!" section mentions that the site provides answers to over 100,000 searches a day. There is another Amazon.com advertisement for "Click to Pay" with the text "Hello Support this site today!". A "Crossword Clue Search" section has a form with "Clue:" and "Pattern:" input fields and a "Go!" button. Below the form, there is a "How to Search:" section explaining the search process. At the bottom, there is a table of example clues and patterns.

ONE
ACROSS

Home [Crosswords](#) [Cryptograms](#) [Anagrams](#) [Reference](#) [Forum](#) [Languages](#)
News! [Contact@](#) [Support](#) [Advertise](#) [Privacy](#) [About](#)

SHOP FOR **new and used** amazon.com and you're done!
DVDs

OneAcross Changes
OneAcross is [moving](#) to a bigger better server. This may cause some problems as we get kinks worked out, but hopefully by early next week, everything will be in better shape. Thanks for your patience!

Support this Site!
We currently provide answers to over 100,000 searches a day. You can help!

Hello Support this site today!
Click to Pay fully refundable
amazon honor system

Crossword Clue Search
Having trouble getting the last word in that puzzle?
Having trouble getting the first? See if our search engine can help! Unlike pure pattern dictionary searches, we actually analyze the clue as well.

Clue:
Pattern:

How to Search: Enter a clue and either the length of the answer or an answer pattern. For unknown letters in the word pattern, you can use a question mark.

Clue: Trout Basket	Clue: Cut
Pattern: 5	Pattern: ???n
Clue: ?	Clue: Scheme
Pattern: ?a?T??s??ke	Pattern: F...

[Hints for better searching...](#)

Example Applications of NLP: Question Answering

AnswerBus Question Answering System - who is married to bill gates

http://www.answerbus.com/cgi-bin/answerbus/answer

Search: microsoft spelling correction

AnswerBus

who is married to bill gates?

Type in your question in English, French, Spanish, German, Italian or Portuguese.

Question:

who is married to bill gates

Possible answers: [XML](#) [TXT](#)

- [Bill was married to Melinda French Gates in 1994 in Hawaii.](#)
- [Mary Gates, Bill's mother, biggest fan, and strongest prodder, finally laid down an ultimatum in 1993. She was dying of cancer, and wanted to see her only son married.](#)
- [Bill Gates married Melinda French in Hawaii on January 1, 1994, and his mother died a few months later.](#)
- [1994 Bill Gates and Melinda French married in Hawaii on New Years Day.](#)

Try your question on other engines:

[Alta Vista](#) | [CNN News Engine](#) | [Ask Jeeves](#) | [Excite](#) | [Google](#) | [HotBot](#) | [Lycos](#) | [Start](#) | [Yahoo](#)

Example Applications of NLP: Machine Translation

amazon.de | WUNSCHZETTEL | MEIN KONTO | HILFE

HOME | MEIN SHOP | **BÜCHER** | ENGLISH BOOKS | ELEKTRONIK & FOTO | KÜCHE & HAUSHALT | MUSIK | DVD | VIDEO | SOFTWARE | COMPUTER & VIDEOSPIELE

EXTENDED SEARCH | STOEBERN | BEST-SELLER | NOVELTIES | SPECIALIZED BOOKS | TIME WRITINGS | PRICE HIT | USED

High-speed search: German books | | Stoebern: All categories |

LOS

ALLES MUSS RAUS - JETZT ZUGREIFEN! Solange der Vorrat reicht | Schnäppchen jagen

BUCH-INFO

More to this book

[Overview](#)

[Amazon.de reader](#)

More of...


[Gerhard Baumfalk](#)

What do you mean?

[Their opinion to this book](#)

[Continue to recommend the book by E-Mail](#)

Assault or preventive strike? The German attack on the Soviet Union on 22 June 1941.
of [Gerhard Baumfalk](#)



Used & again starting from **EUR 10,50**
Offerer dispatches in 1-2 working-days.

ALLE ANGEBOTE

1 uses 10,50 starting from EUR

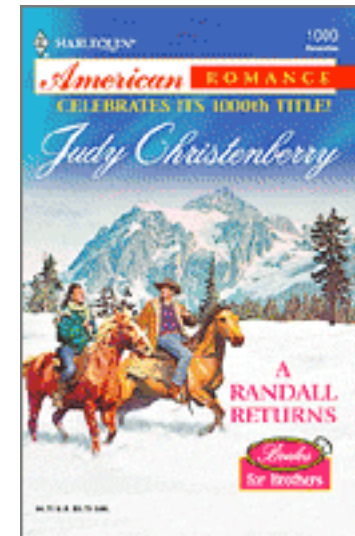
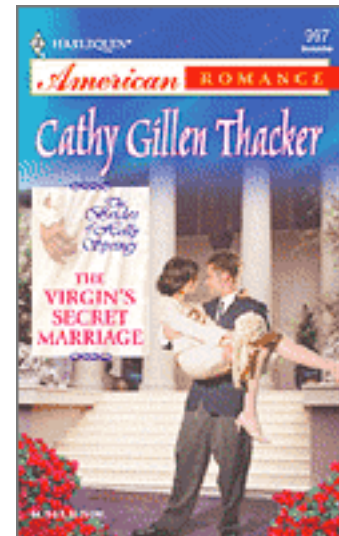
Would you like to sell?
[Diesen Artikel verkaufen](#)

Kategorie(n): [Politics, Biografien & history](#), [specialized books](#)

Now get money back!
Now fetch the taxes back, which are entitled to you -- with the software and

Paperback - 164 sides - Rita Fischer, Ffm.
Publication date: July 1997
ISBN: 3895014931
Amazon.de Verkaufsrang 204.896

Example Applications of NLP: Automatically generate Harlequin Romance novels?



Goals of the Course

- Introduce you to Natural Language Processing problems and solutions.
- Ultimate focus on handling ambiguity by probabilistic integration of evidence.
- Give you some hands-on practice with data and a handful of methods.

This Class

- Assumes you come with some skills...
 - Some basic math/probability, decent programming skills
(We will use Python; tutorial coming next week.)
 - Some ability to learn missing knowledge
- Teaches key theory and methods for language modeling, tagging, parsing, etc.
- But it's something like an “AI Systems” class:
 - Hands on with data
 - Often practical issues dominate over theoretical niceties

Course Logistics

- Professor: Andrew McCallum
- TAs: David Mimno
Karl Schultz
- Assistants: Hanna Wallach
Khash Rohanimanesh
- Time: Tue/Thu 2:30-3:45pm
- Mailing list: `585-staff@cs.umass.edu`
- More information on Web site:
<http://www.cs.umass.edu/~mccallum/courses/inlp2007>

Grading

- 7 short written homework / programming assignments.
 - no way to really internalize without doing it
 - some hands-on experience
 - should be fun!
 - should take about 1-2 hours each.
- Random, informal in-class “collaborative quizzes”
 - help you set expectations for the mid-term and final
- Final project: with a small team, mixed backgrounds
 - chance to explore a special interest at end of term
- Midterm & Final, and classroom participation

For Linguistics Students: Programming? Yipes!

- Yes, but with *extensive* support for those w/out experience.
- Historically popular language for CL courses:
 - Prolog (clean, hard to learn, counter-intuitive)
 - Perl (quick, but obfuscated syntax, messy to read)
 - Interpreted, rapid prototyping
- Why **Python** is better-suited:
 - easy to learn, clean syntax, powerful features
 - becoming increasingly popular in CompLinguistics!
 - Extensive tutorials, CompLing support, toolkits, data, etc.
- Many CS students don't know it either: put you on more equal footing.

Syllabus Outline

- Two parts:
 - First: hands-on course, introductory, methods, HW
 - Second: more like a seminar + project
- First half:
 - Language, structures, and computation
 - Foundation of probability and information theory
 - Use those foundations to work with language
- Example topics:
 - Language models, language prediction, spam filtering.
 - Collocations, word clustering, word sense disambiguation.
 - Finite state machines, Markov models, Part-of-speech tagging.
 - Modern parsing techniques.
 - Information extraction, semantics, question answering, discourse.

To Do This Week

- Visit course Web site, browse around.
- Read Chapters 1 and 2 in Jurafsky & Martin textbook
 - Available on line! See course web site.
- Install Python on your computer
 - Get extensive help from the TAs if you like!

Thank you!