Algorithms for NLP

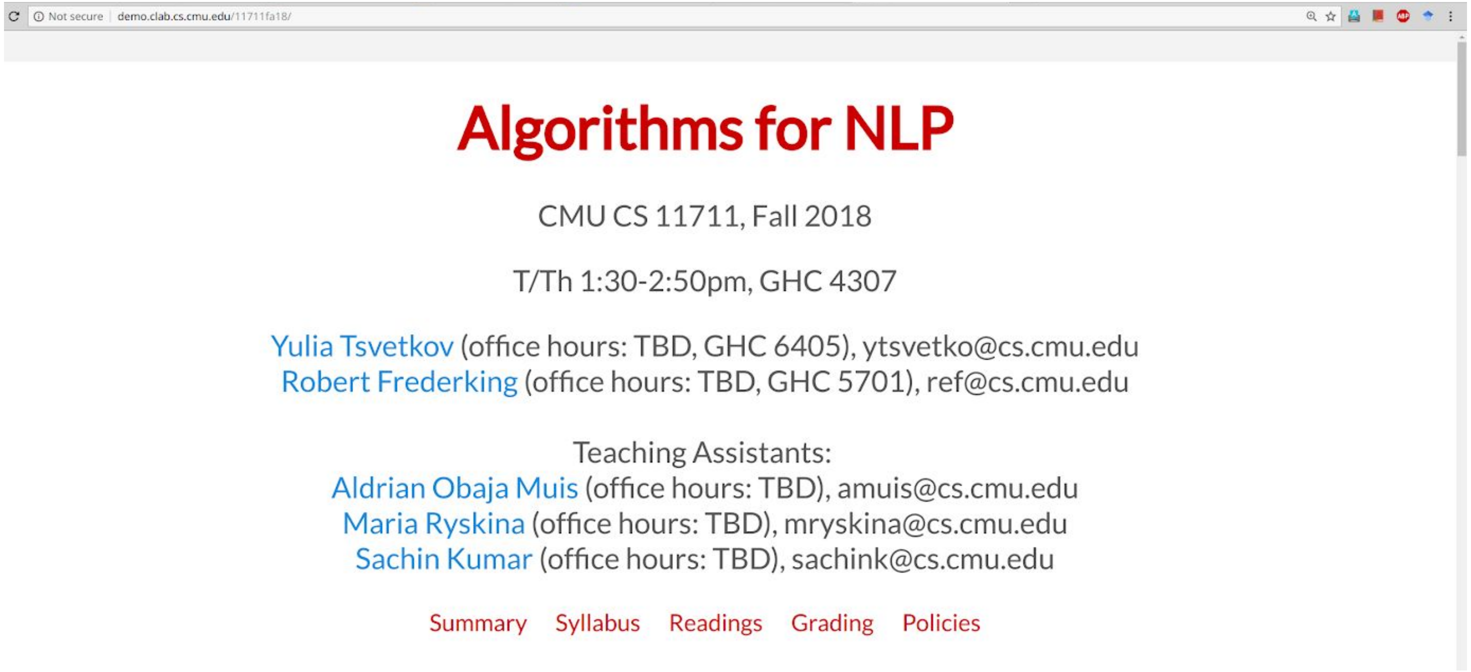
Lecture 1: Introduction

Yulia Tsvetkov – CMU

Slides: Nathan Schneider – Georgetown,
Taylor Berg-Kirkpatrick – CMU/UCSD,
Dan Klein, David Bamman – UC Berkeley

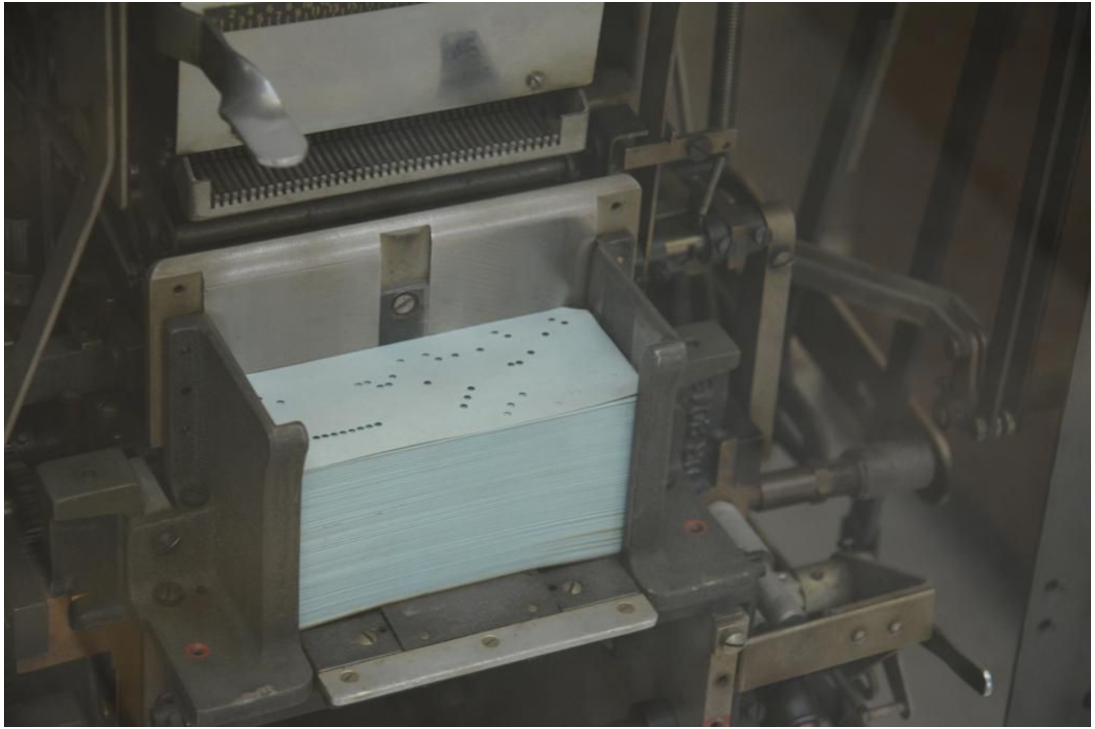
Course Website

http://demo.clab.cs.cmu.edu/11711fa18/



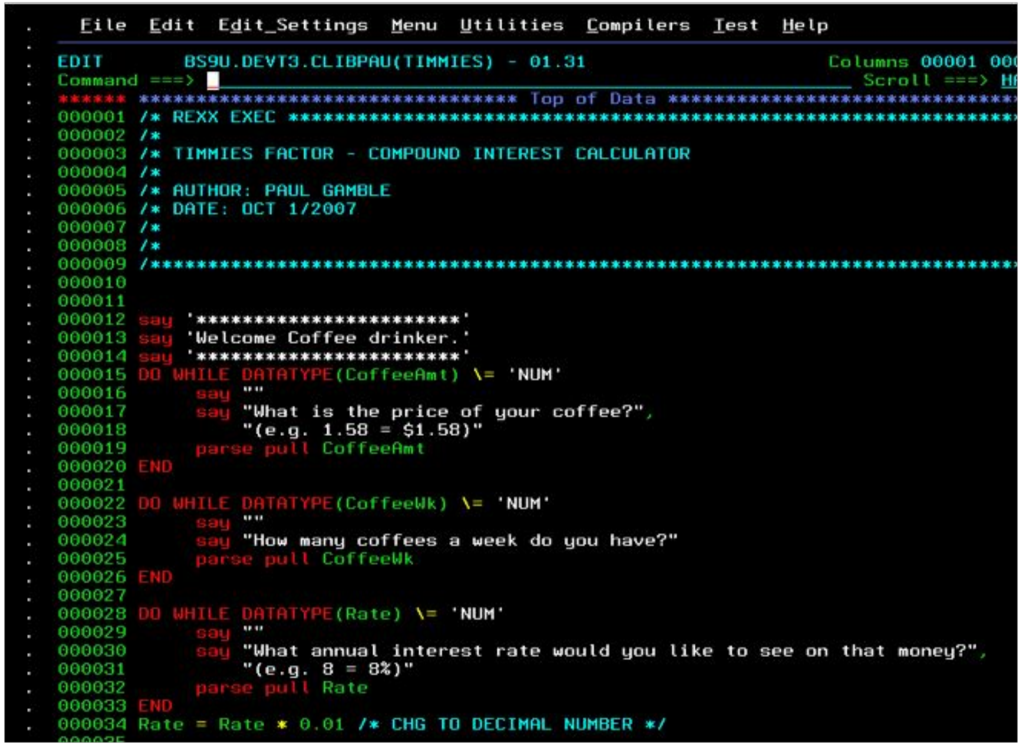
Communication with Machines

- ~50s-70s



Communication with Machines

- ~80s



Communication with Machines

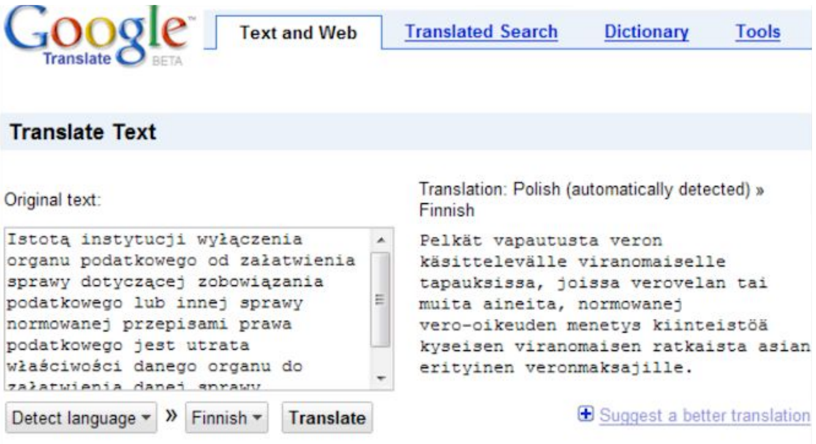
- Today

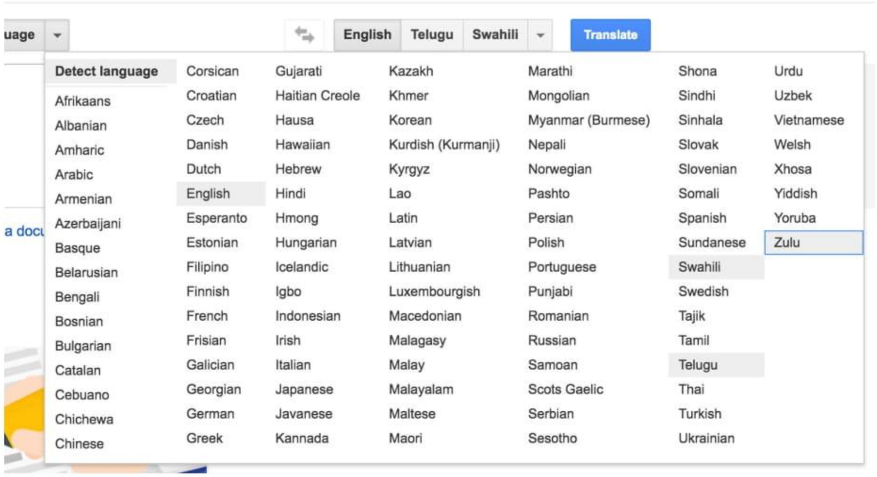
Language Technologies

WeKnowMemes

- A conversational agent contains
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Information retrieval
  - Text to speech

Language Technologies





Language Technologies

- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England in the end of the 18th century?
- What do scientists think about the ethics of human cloning?

Natural Language Processing

- Applications
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - ...
- Core technologies
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Coreference resolution
  - Word sense disambiguation
  - Semantic Role Labelling
  - ...

NLP lies at the intersection of **computational linguistics** and **artificial intelligence**. NLP is (to various degrees) informed by linguistics, but with practical/engineering rather than purely scientific aims.

What does an NLP system need to ‘know’?

- Language consists of many levels of structure
  - Humans fluently integrate all of these in producing/understanding language
  - Ideally, so would a computer!

Phonology

SOUNDS Th i a si e n

- Pronunciation modeling

Example by Nathan Schneider

Words

WORDS This is a simple sentence

- Language modeling
- Tokenization
- Spelling correction

Example by Nathan Schneider

Morphology

WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

- Morphological analysis
- Tokenization
- Lemmatization

Example by Nathan Schneider

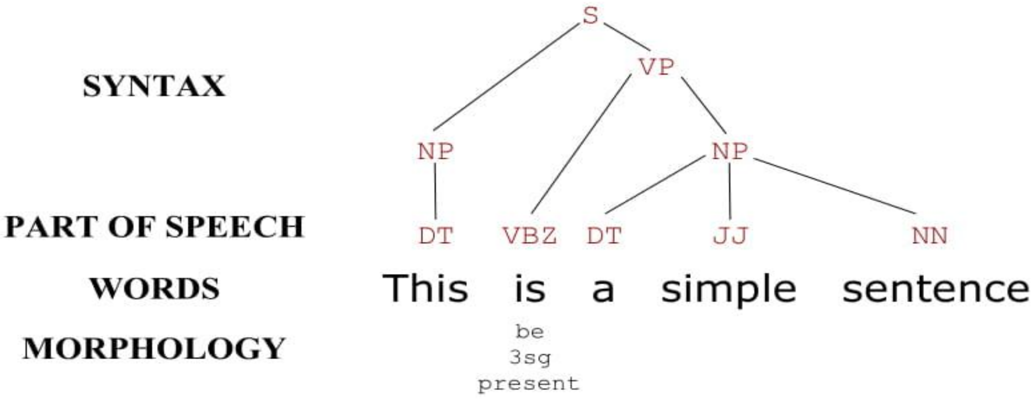
Parts of speech

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

- Part-of-speech tagging

Example by Nathan Schneider

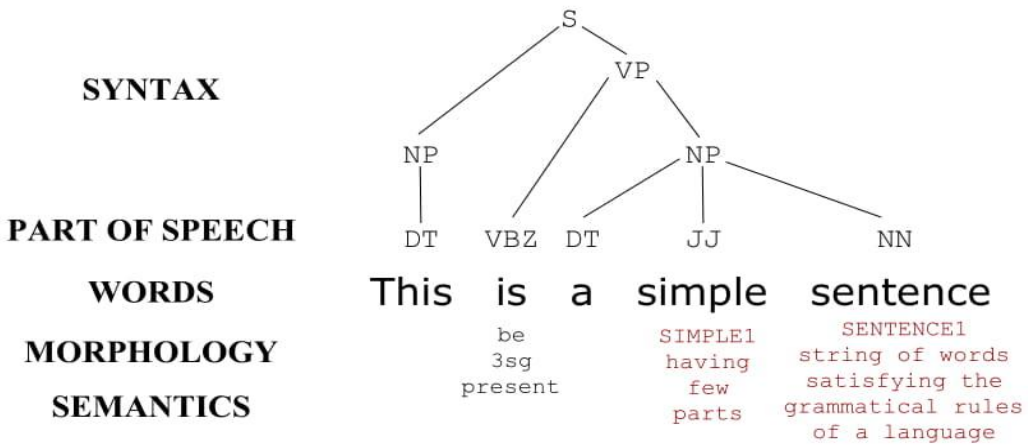
Syntax



- Syntactic parsing

Example by Nathan Schneider

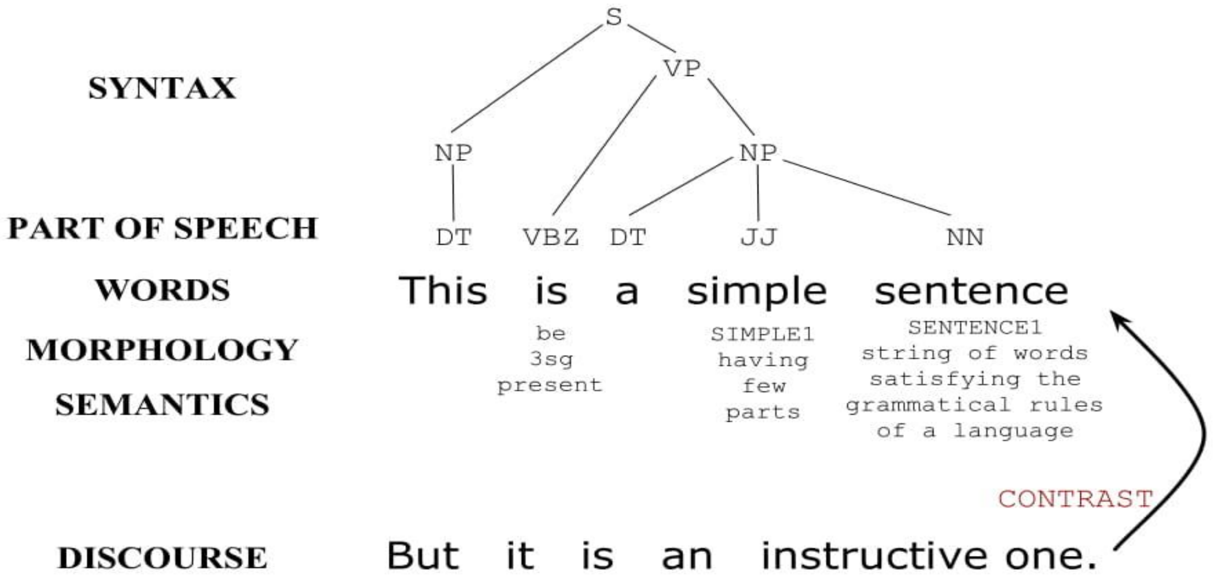
Semantics



- Named entity recognition
- Word sense disambiguation
- Semantic role labelling

Example by Nathan Schneider

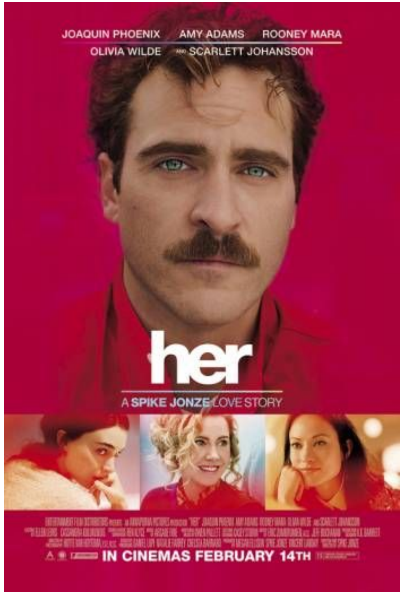
Discourse



- Reference resolution

Example by Nathan Schneider

Where We Are Now?



Baseline mutual information model (Li et al. 2015)
A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...
A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about . (6)
A: You don't know what you are saying. (7)
...

Li et al. (2016), "Deep Reinforcement Learning for Dialogue Generation" *EMNLP*

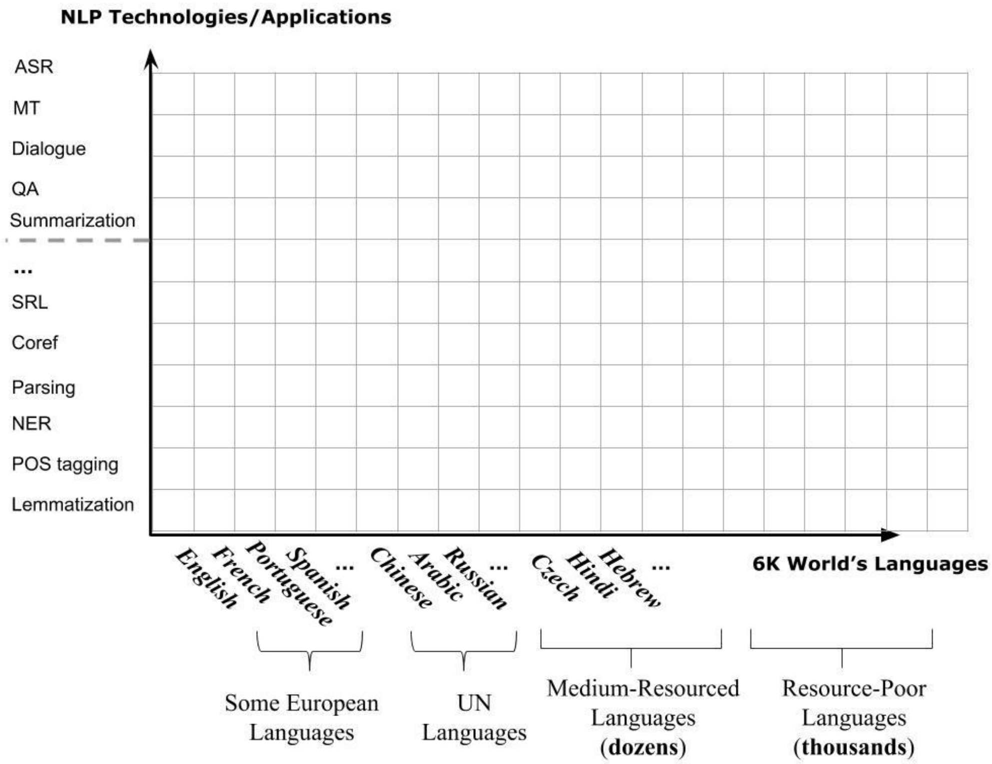
Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation

Ambiguity

- Ambiguity at multiple levels:
  - Word senses: **bank** (finance or river?)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I saw her duck**

Scale + Ambiguity



Tokenization

这是一个简单的句子

WORDS

This is a simple sentence

זה משפט פשוט

Word Sense Disambiguation

in tea
her daughter

בתה

- most of the vowels unspecified

Tokenization + Disambiguation

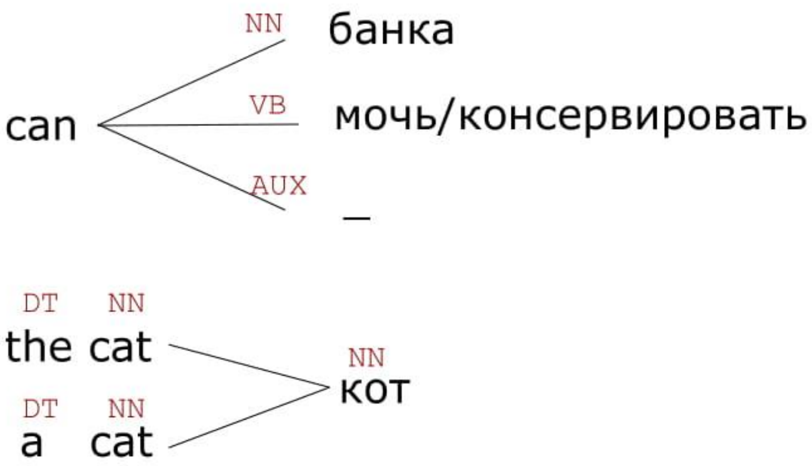
in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Part of Speech Tagging



Tokenization + Morphological Analysis

- Quechua morphology

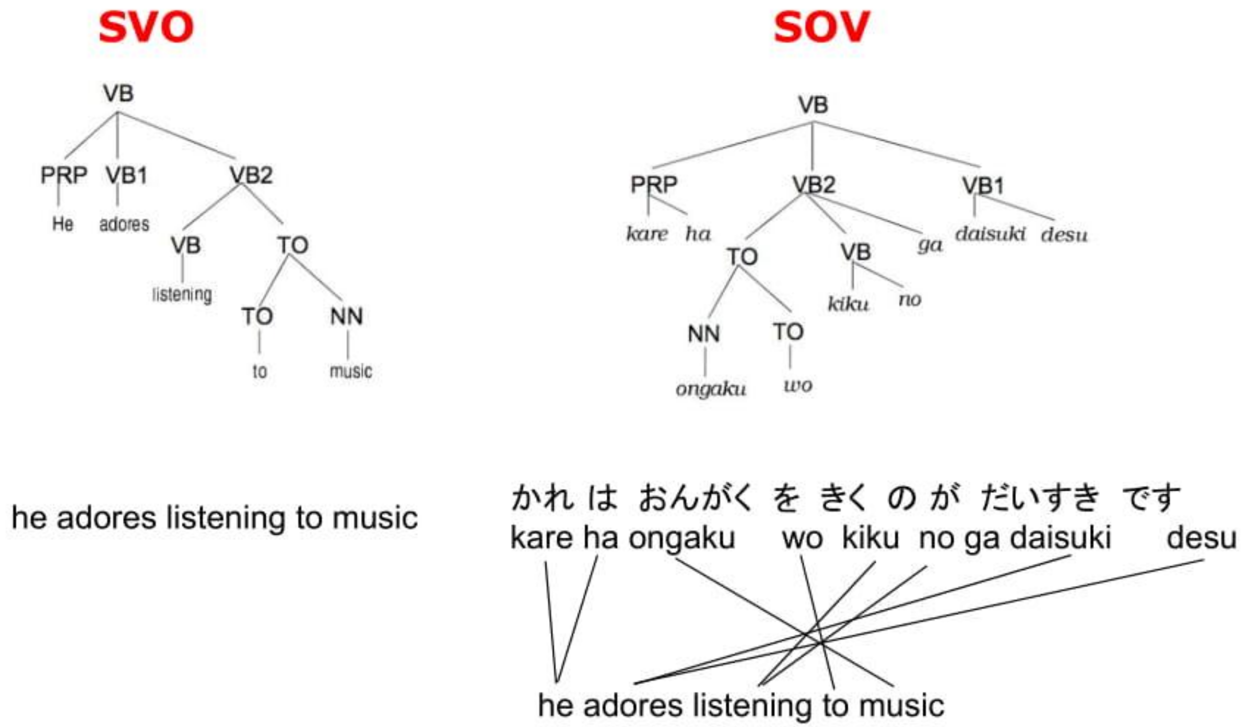
Much’ananayakapushasqakupuniñataqsunamá

Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“So they really always have been kissing each other then”

```
Much’a  to kiss
-na     expresses obligation, lost in translation
-naya   expresses desire
-ka     diminutive
-pu     reflexive (kiss *eachother*)
-sha    progressive (kiss*ing*)
-sqa    declaring something the speaker has not personally witnessed
-ku     3rd person plural (they kiss)
-puni   definitive (really*)
-ña     always
-taq    statement of contrast (...then)
-suna   expressing uncertainty (So...)
-má     expressing that the speaker is surprised
```

Syntactic Parsing, Word Alignment



Semantic Analysis

- Every language sees the world in a different way
  - For example, it could depend on cultural or historical conditions

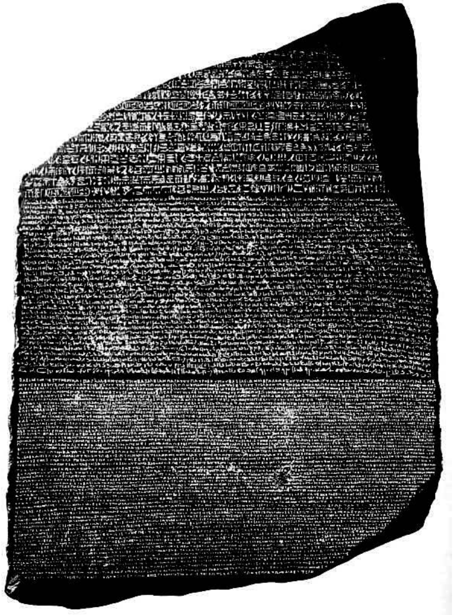
  - Russian has very few words for colors, Japanese has hundreds
  - Multiword expressions, e.g. it's raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages

Dealing with Ambiguity

- How can we model ambiguity and choose the correct analysis in context?
  - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return *all possible analyses*.
  - probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis,* i.e., the most probable one according to the model.

- But the "best" analysis is only good if our probabilities are accurate. Where do they come from?

Corpora



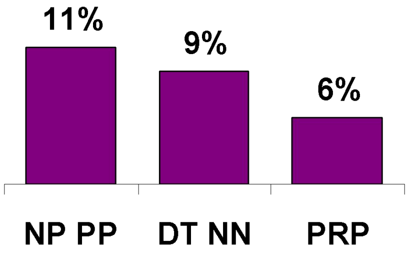
- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text

- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - Yelp reviews
  - The Web: billions of words of who knows what

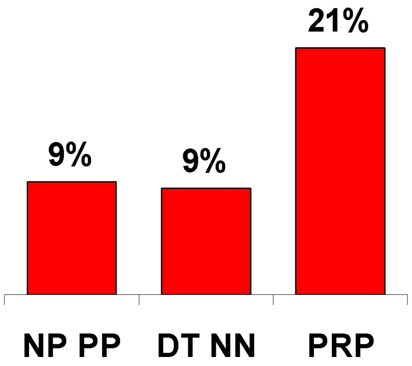
Corpus-Based Methods

- Give us statistical information

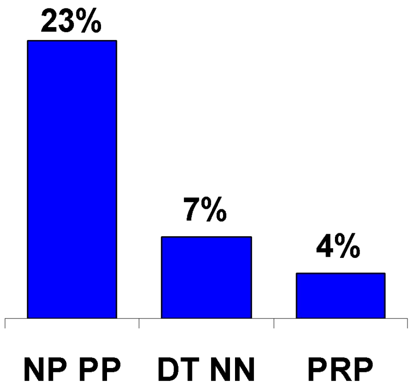
All NPs



NPs under S

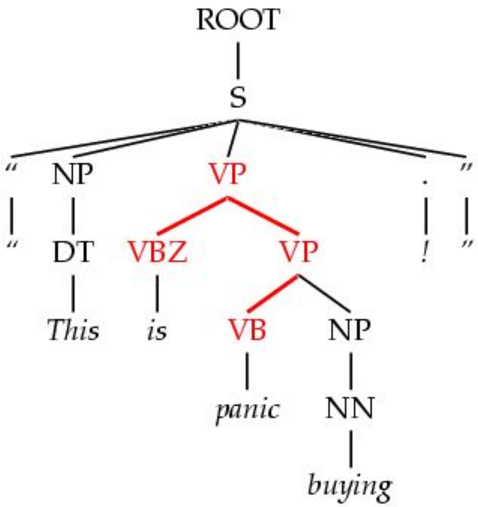


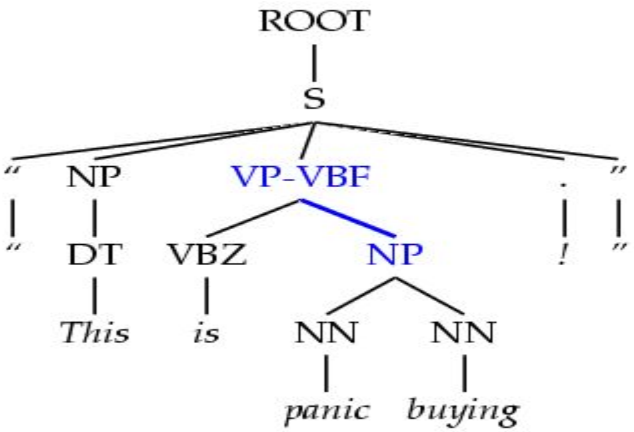
NPs under VP



Corpus-Based Methods

- Let us check our answers





TRAINING	DEV	TEST

Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
  - Typically more robust than earlier rule-based methods
  - Relevant statistics/probabilities are *learned from data*
  - Normally requires *lots of data* about any particular phenomenon

Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation

Sparsity

- Sparse data due to Zipf's Law
  - To illustrate, let's look at the frequencies of different words in a large text corpus
  - Assume "word" is a string of letters separated by spaces

Word Counts

Most frequent words in the English Europarl corpus (out of 24m word **tokens**)

any word

Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns

Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Word Counts

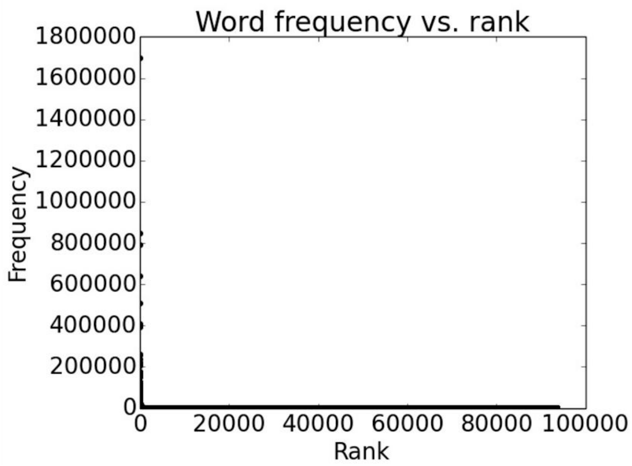
But also, out of 93,638 distinct words (**word types**), 36,231 occur only once.

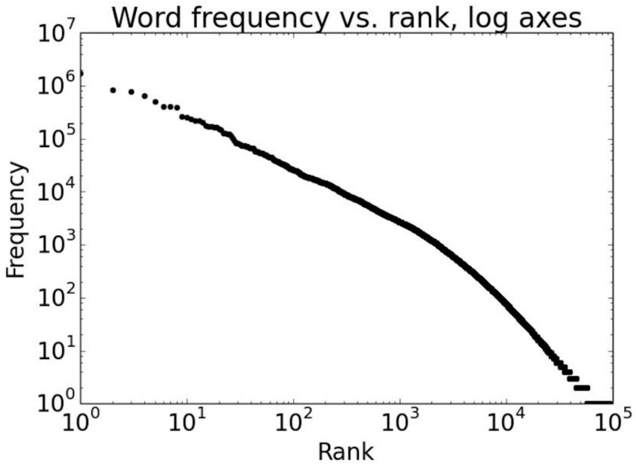
Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

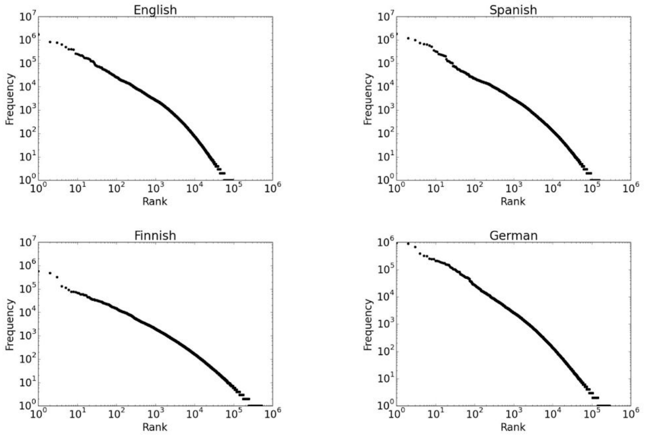
Plotting word frequencies

Order words by frequency. What is the frequency of n_{th} ranked word?





Zipf's Law



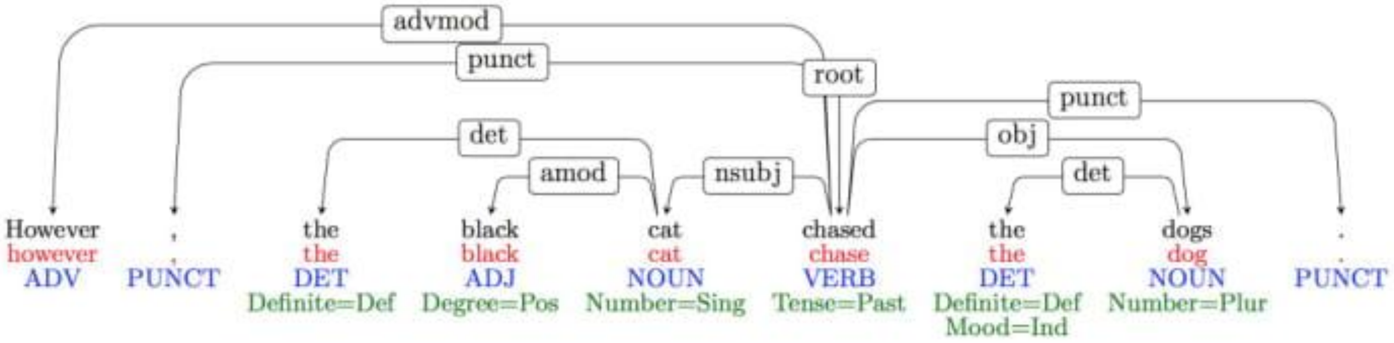
- Implications
  - Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
  - This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

Why is NLP Hard?

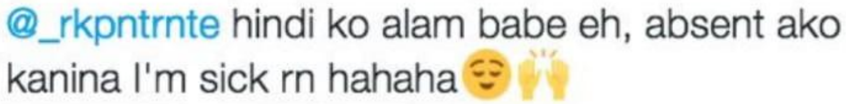
1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation

Variation

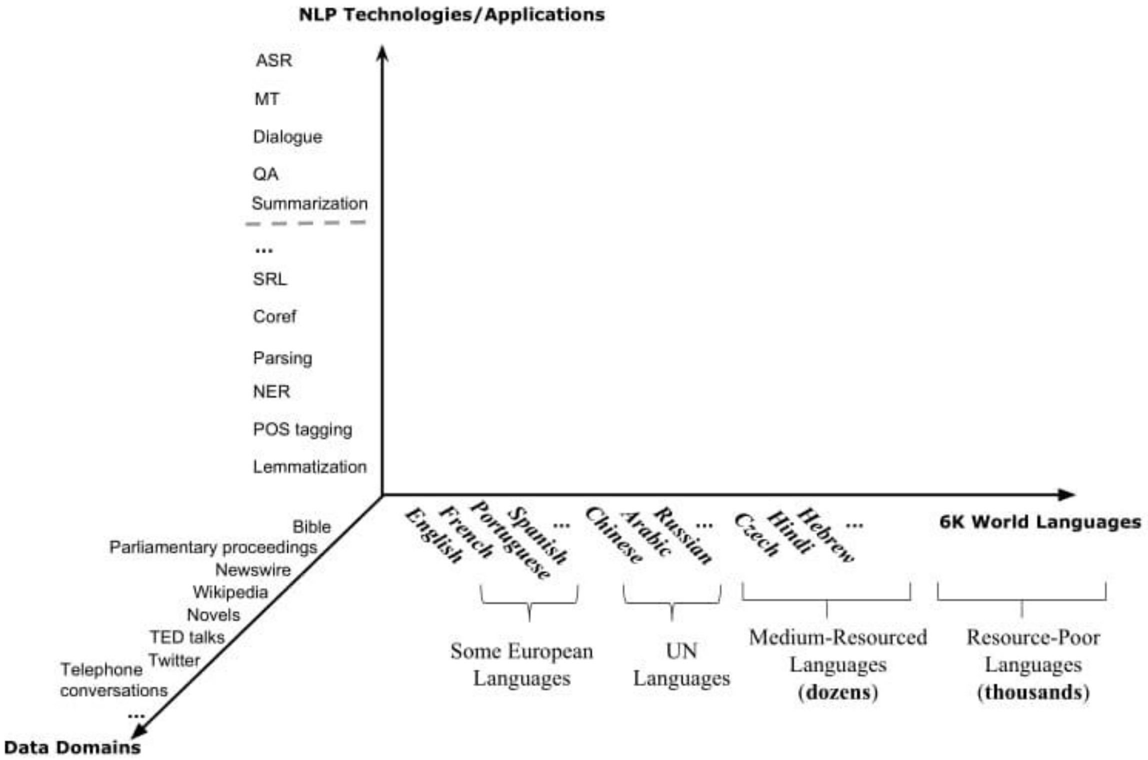
- Suppose we train a part of speech tagger or a parser on the Wall Street Journal



- What will happen if we try to use this tagger/parser for social media??



Why is NLP Hard?



Why is NLP Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation

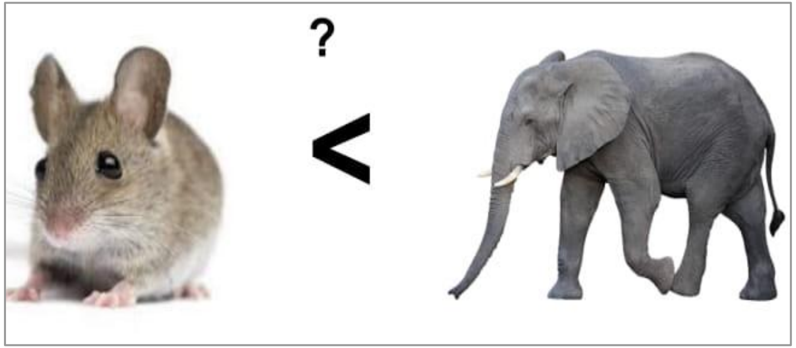
Expressivity

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

  - She gave the book to Tom vs. She gave Tom the book
  - Some kids popped by vs. A few children visited
  - Is that window still open? vs. Please close the window

Unmodeled variables

“Drink this milk”



- World knowledge
  - I dropped the glass on the floor and it broke
  - I dropped the hammer on the glass and it broke

Unknown Representation

- Very difficult to capture, since we don't even know how to represent the knowledge a human has/needs: What is the "meaning" of a word or sentence? How to model context? Other general knowledge?

Models and Algorithms

- Models
  - State machines (finite state automata/transducers)
  - Rule-based systems (regular grammars, CFG, feature-augmented grammars)
  - Logic (first-order logic)
  - Probabilistic models (WFST, language models, HMM, SVM, CRF, ...)
  - Vector-space models (embeddings, seq2seq)
- Algorithms
  - State space search (DFS, BFS, A*, dynamic programming---Viterbi, CKY)
  - Supervised learning
  - Unsupervised learning
- Methodological tools
  - training/test sets
  - cross-validation

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…

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 Applications
  - Reference resolution
  - Summarization
  - …

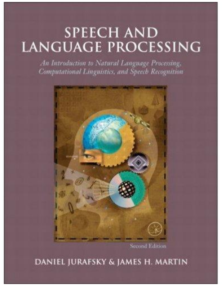
Requirements and Goals

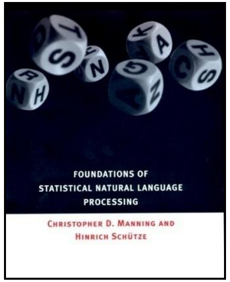
- Class requirements
  - Uses a variety of skills / knowledge:
    - Probability and statistics, graphical models
    - Basic linguistics background
    - Strong coding skills (Java)
  - Most people are probably missing one of the above
  - You will often have to work on your own to fill the gaps

- 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!

Logistics

- Prerequisites:
  - 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, 2nd and 3rd Edition (not 1st)
  - Also: Manning and Schuetze, Foundations of Statistical NLP





Other Announcements

- Course Contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
  - Canvas: project submissions
  - Homework questions: Recitations, Piazza, TAs’ office hours
- 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?

Some Early NLP History

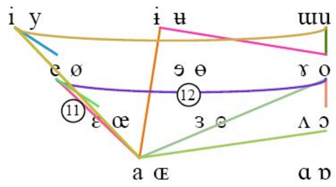
- 1950's:
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!
- 1960's and 1970's: NLP Winter
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - 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*

A More Recent NLP History

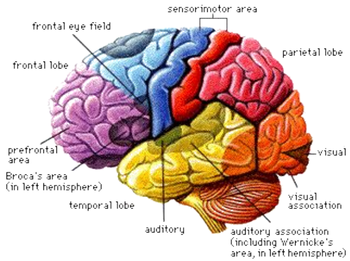
- 2000+: Richer Statistical Methods
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - *Begin to get both breadth and depth*
- 2013+: Deep Learning

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'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

http://demo.clab.cs.cmu.edu/11711fa18/