I256:
Applied Natural Language Processing

Marti Hearst
Week 4

From the Homework: Compare These Two Collections

- know 1.58%
- right 1.49%
- what 1.31%
- s**t 1.16%
- like 1.02%
- yeah 0.97%
- f**k 0.85%
- that 0.83%
- come 0.67%
- back 0.66%
- need 0.66%
- good 0.62%
- want 0.62%
- they 0.53%
- think 0.49%
- f**king 0.46%

```
Total No. of sentences: 53271
Avg. sent. len (chars): 15.34
Avg. sent. len (words): 2.73
```

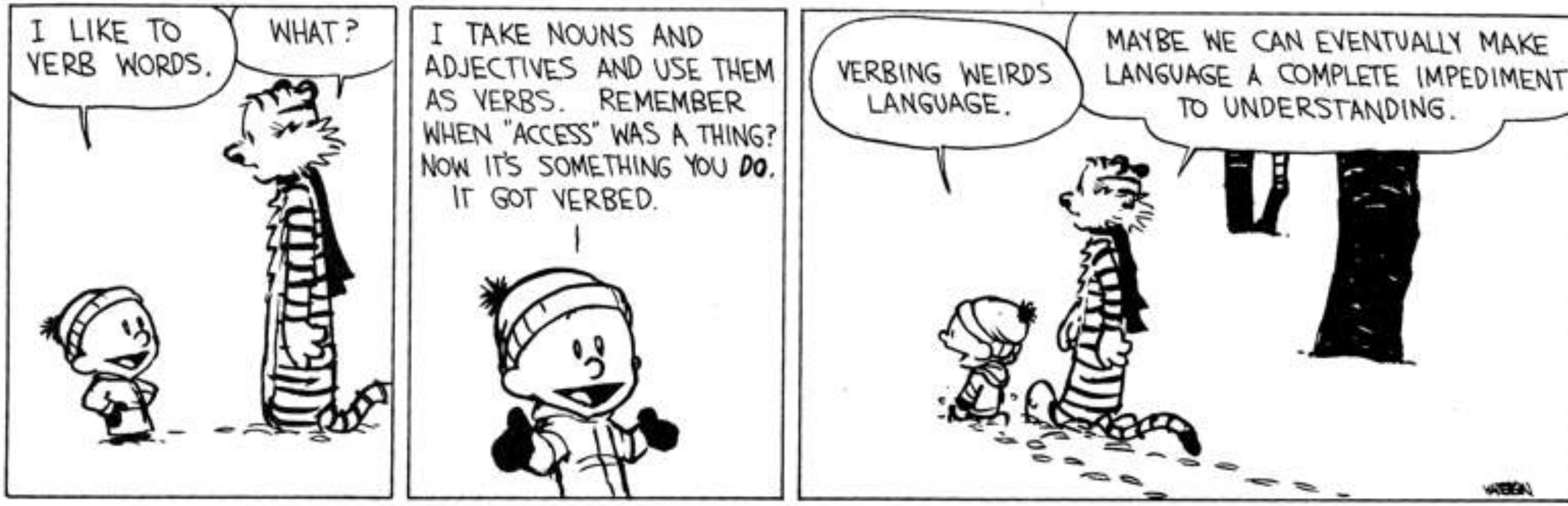
- people 1.47
- patent 1.11
- really 0.73
- Ventures 0.59
- Intellectual 0.59
- actually 0.58
- didn't 0.54
- company 0.54
- patents 0.53
- something 0.52
- things 0.51
- called 0.5
- that's 0.49
- companies 0.49
- you're 0.48
- litt
- Th
- nu

Info type	Value
Number of Sentences	6778
Average Length in Words	15
Average Length in Characters	72

Some Interesting Results for Capitalized Words

- ('Earth', 246),
- ('Mars', 76),
- ('Galileo', 75),
- ('Herschel', 70),
- ("Earth's", 70),
- ('American', 68),
- ('Jupiter', 67),
- ('Hale', 62),
- ('Andromeda', 57),
- ('Saturn', 50),
- ('Venus', 49),
- ('John', 46),
- ('Since', 45),
- ('William', 43),
- ('Observatory', 43),
- ('Milky', 42),
- ('Sirius', 42),

- Ser (0.32%)
- | Lord (0.30%)
- | Jon (0.28%)
- | Ned (0.27%)
- | Tyrion (0.21%)
- | Bran (0.18%) | Catelyn (0.17%) | Arya (0.16%) | Sansa (0.14%) | Dany (0.14%) | Robb (0.14%) | Robert (0.14%) | Stark (0.13%) | Lannister (0.12%) | Maester (0.09%) | King (0.08%) | Winterfell (0.08%) | Joffrey (0.08%) | Drogo (0.07%) | Eddard (0.06%) | Jorah (0.06%) | Lady (0.06%) | Littlefinger (0.05%) | Hand (0.05%) | Mormont (0.05%) | Tywin (0.05%) | Dothraki (0.05%) | Jaime (0.05%) | Wall (0.05%) | Luwin (0.05%)



Source: Calvin and Hobbs

Today

- What are parts of speech (POS)?
- Practice with parts of speech identification.
- Intro to N-grams / Language Models

COURSE CONCEPT MAP

	Word-Level	**Phrase-Level**	**Sentence-Level**	**Discourse-Level**
Segment-ation				
Syntax	Part of Speech Tagging			
Semant-ics				

Statistics

Why Parts of Speech?

- A word's POS says a lot about the word **and its neighbors**:
  - Limits the range of meanings (*deal v vs n*), pronunciation, (*object vs object*), or both (*wind v vs n*)
  - Helps in stemming:

    saw[v] → see,

    saw[n] → saw
  - Can help select nouns for summarization
  - Useful for information extraction
  - Helpful for semantic analysis
    - Noun-noun compounds have important meaning

How do we Define Parts of Speech?

- By meaning
  - Verbs are actions
  - Adjectives are properties
  - Nouns are things
- By the syntactic environment
  - What occurs nearby?
  - What does it act as?
- By what morphological processes affect it
  - What affixes does it take?
- Combination of the above

Two Types of Parts of Speech

- Closed class
  - Closed, fixed membership
  - Reasonably easy to enumerate
  - Generally, short function words that "structure" sentences
- Open class
  - Impossible to completely enumerate
  - New words continuously being invented, borrowed, etc.

Nouns

- Open class
  - New inventions all the time: *muggle*, *webinar*, ...
- Semantics:
  - Generally, words for people, places, things
  - But not always (bandwidth, energy, ...)
- Syntactic environment:
  - Occur with determiners
  - Pluralizable, possessivizable
- Other characteristics:
  - Mass vs. count nouns
    - Mass nouns: mud, furniture, taste

Verbs

- Open class
  - New inventions all the time: google, tweet, ...
- Semantics:
  - Generally, denote actions, processes, etc.
- Syntactic environment:
  - Intransitive, transitive, ditransitive
  - Alternations
    - Jane broke the window. vs. The window broke.
- Other characteristics:
  - Main vs. auxiliary verbs
  - Gerunds (verbs behaving like nouns)
    - Reading is a good way to learn.
  - Participles (verbs behaving like adjectives)
    - The rising sun

Closed Class POS

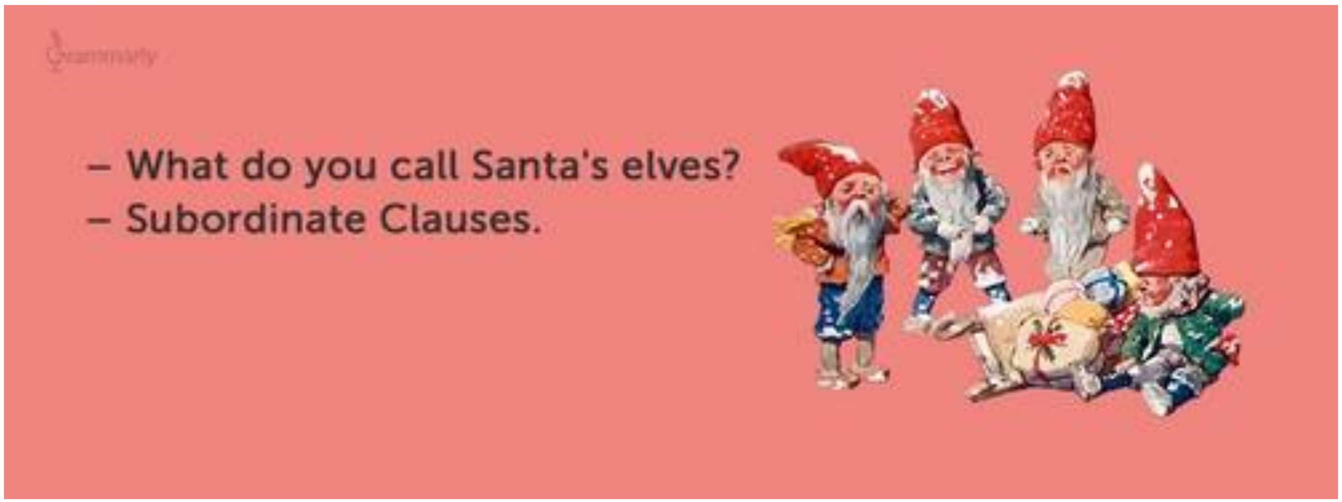
- Prepositions
  - In English, occurring before noun phrases
  - Specifying some type of relation (spatial, temporal, …)
  - Examples: *on* the shelf, *before* noon
- Particles
  - Resembles a preposition, but used with a verb
    - often change the core meaning ("phrasal verbs")
    - find *out*, turn *over*, go *on*

Particle vs. Prepositions Exercise: which is which?

He came *by* the office in a hurry (by = preposition)
He came *by* his fortune honestly (by = particle)

We ran *up* the phone bill (up = particle)
We ran *up* the small hill (up = preposition)

He lived *down* the block (down = preposition)
He never lived *down* the nicknames (down = particle)


– What do you call Santa's elves?
– Subordinate Clauses.

Closed Class POS: Conjunctions

- Coordinating conjunctions
  - Join two elements of "equal status"
    - Examples: cats *and* dogs, salad *or* soup
- Subordinating conjunctions
  - Join two elements of "unequal status"
  - Examples:
    - We'll leave *after* you finish eating.
    - *While* I was waiting in line, I saw my friend.
  - Complementizers are a special case:
    - I think *that* you should finish your assignment

Penn Treebank Tagset: 45 Tags

- Traditional grammar classifies words based on eight parts of speech:
  - verb (VB)
  - noun (NN)
  - pronoun (PR+DT)
  - adjective (JJ)
  - adverb (RB),
  - preposition (IN),
  - conjunction (CC),
  - interjection (UH)
- Penn Treebank goes into far more detail.
- Manually assigned POS tags to many sentences.

Penn Treebank POS Tags

Tag	Description	Example
CC	conjunction, coordinating	*and, or, but*
CD	cardinal number	*five, three, 13%*
DT	determiner	*the, a, these*
EX	existential there	*there were six boys*
FW	foreign word	*mais*
IN	conjunction, subordinating or preposition	*of, on, before, unless*
JJ	adjective	*nice, easy*
JJR	adjective, comparative	*nicer, easier*
JJS	adjective, superlative	*nicest, easiest*
LS	list item marker	
MD	verb, modal auxillary	*may, should*
NN	noun, singular or mass	*tiger, chair, laughter*
NNS	noun, plural	*tigers, chairs, insects*
NNP	noun, proper singular	*Germany, God, Alice*
NNPS	noun, proper plural	*we met two Christmases ago*
PDT	predeterminer	*both his children*
POS	possessive ending	*'s*
PRP	pronoun, personal	*me, you, it*
PRP$	pronoun, possessive	*my, your, our*
RB	adverb	*extremely, loudly, hard*
RBR	adverb, comparative	*better*

Penn Treebank POS Tags

RBS	adverb, superlative	*best*
RP	adverb, particle	*about, off, up*
SYM	symbol	*%*
TO	infinitival to	*what to do?*
UH	interjection	*oh, oops, gosh*
VB	verb, base form	*think*
VBZ	verb, 3rd person singular present	*she thinks*
VBP	verb, non-3rd person singular present	*I think*
VBD	verb, past tense	*they thought*
VBN	verb, past participle	*a sunken ship*
VBG	verb, gerund or present participle	*thinking is fun*
WDT	*wh*-determiner	*which, whatever, whichever*
WP	*wh*-pronoun, personal	*what, who, whom*
WP$	*wh*-pronoun, possessive	*whose, whosever*
WRB	*wh*-adverb	*where, when*
.	punctuation mark, sentence closer	*.;?**
,	punctuation mark, comma	*,*
:	punctuation mark, colon	*:*
(	contextual separator, left paren	*(*
)	contextual separator, right paren	*)*

EXERCISE

POS INTERPRETATION

POS Tag Practice: Which Tags are Incorrect?

The Part-of-Speech tagger has automatically labeled the input in the following way.

PRP/ I MD/ would NN/ fain NN/ bestow CC/ and VB/ distribute ,/ , IN/ until DT/ the JJ/ wise VBP/ have RB/ once JJR/ more VBN/ become
JJ/ joyous IN/ in PRP$/ their NN/ folly ,/ , CC/ and DT/ the JJ/ poor JJ/ happy IN/ in PRP$/ their NNS/ riches ./ .

RB/ Therefore MD/ must PRP/ I VBP/ descend IN/ into DT/ the JJ/ deep :/ : IN/ as NN/ thou VBP/ doest IN/ in DT/ the NN/ evening ,/ , WRB/ when NN/ thou NN/ goest IN/ behind DT/ the NN/ sea ,/ , CC/ and JJS/ givest NN/ light RB/ also TO/ to DT/ the
NN/ nether-world ,/ , NN/ thou JJ/ exuberant NN/ star ./ !

Tool at: http://cogcomp.cs.illinois.edu/page/demo_view/POS
Text from Thus Spake Zarathustra by Friedrich Nietzsche

POS Tag Practice:
Which Tags are Incorrect?

The Part-of-Speech tagger has automatically labeled the input in the following way.

IN/ On NNP/ January CD/ 17 ,/ , CD/ 2012 ,/ , DT/ the NNP/ AIRC VBD/ approved JJ/ final JJ/ congressional CC/ and NN/ state JJ/ legislative NNS/ maps VBN/ based IN/ on DT/ the CD/ 2010 NN/ census ./ . VB/ See NNP/ Arizona NNP/ Independent NNP/ Redistricting ,/ , JJ/ Final NNP/ Maps ,/ , NN/ http://azredistricting.org/Maps/Final-Maps/default.asp -LRB-/ (DT/ all NN/ Internet NNS/ materials RB/ as VBN/ visited NNP/ June CD/ 25 ,/ , CD/ 2015 ,/ , CC/ and VBD/ included IN/ in NNP/ Clerk IN/ of NNP/ Court NN/ ' VBZ/ s NN/ case NN/ file -RRB-/) ./ . RBR/ Less IN/ than CD/ four NNS/ months RB/ later ,/ , IN/ on NNP/ June CD/ 6 ,/ , CD/ 2012 ,/ , DT/ the NNP/ Arizona NNP/ Legislature VBD/ filed NN/ suit IN/ in DT/ the NNP/ United NNPS/ States NNP/ District NNP/ Court IN/ for DT/ the NNP/ District IN/ of NNP/ Arizona ,/ , VBG/ naming IN/ as NNS/ defendants DT/ the NNP/ AIRC ,/ , PRP$/ its CD/ five NNS/ members ,/ , CC/ and DT/ the NNP/ Arizona NNP/ Secretary IN/ of NNP/ State ./ . DT/ The NNP/ Legislature VBD/ sought DT/ both DT/ a NN/ declaration IN/ that NNP/ Proposition CD/ 106 CC/ and JJ/ congressional NNS/ maps VBN/ adopted IN/ by DT/ the NNP/ AIRC VBP/ are JJ/ unconstitutional ,/ , CC/ and ,/ , IN/ as JJ/ affirmative NN/ relief ,/ , DT/ an NN/ injunction IN/ against NN/ use IN/ of NNP/ AIRC NNS/ maps IN/ for DT/ any JJ/ congressional NN/ election IN/ after DT/ the CD/ 2012 JJ/ general NN/ election ./ .

Tool at: http://cogcomp.cs.illinois.edu/page/demo_view/POS Text from US Supreme Court Opinion by Justice Ginsburg, Ariz. State Legislature v. Ariz. Indep. Redistricting Comm'n 576 U.S. ___ (2015)

POS Tagging: What's the task?

- Process of assigning part-of-speech tags to words
- But what tags are we going to assign?
  - Coarse grained: noun, verb, adjective, adverb, …
  - Fine grained: {proper, common} noun
  - Even finer-grained: {proper, common} noun ± animate
- Important issues to remember
  - Choice of tags encodes certain distinctions/non-distinctions
  - Tagsets will differ across languages!
- For English, Penn Treebank is the most common tagset

Why is it hard?

Number of words that have the corresponding number of tags.

		87-tag Original Brown		45-tag Treebank Brown	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

(Brief Intro)

NGRAMS AND LANGUAGE MODELS

Language Models: Models of **likely word sequences**

- Pay attention to the preceding words
  - "Let's go outside and take a [____]"
    - walk: very likely
    - break: quite likely
    - stone: less likely

- Compute conditional probability as:
  - P(walk | let's go outside and take a)

N-Gram Language Models

N=1 (unigrams)

This is a sentence

This,
is,
a,
sentence

Sentence of length s, how many unigrams?

N-Gram Language Models

N=2 (bigrams)

This is a sentence

This is,
is a,
a sentence

Sentence of length s, how many bigrams?

N-Gram Language Models

N=3 (trigrams)

This is a sentence

This is a,
is a sentence

Sentence of length s, how many trigrams?

Why Language Models?

- POS Tagging:
  - P(n follows det) > P(v follows det)
- Spelling Correction
  - The office is about fifteen **minuets** from my house
    - P(about fifteen minutes from) > P(about fifteen minuets from)
- Speech Recognition
  - P(I saw a van) >> P(eyes awe of an)
- Summarization, question answering, etc etc

What is Language Modeling?

- Goal: compute the probability of a sentence or a sequence of words:
  - $P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$
- Related task: probability of an upcoming word:
  - $P(w_5|w_1,w_2,w_3,w_4)$
- A model that computes either of these two:
  - $P(W)$ or $P(w_n|w_1,w_2 \ldots w_{n-1})$
- is called a **language model**.
  - (Jurafsky thinks a better term is **a grammar**, but LM is standard.)

Probability of a Word Sequence

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- Problem?
- Too many possible sentences! Not enough data to make good estimates.
- Solution: approximate the probability of a word given all the previous words.

Tomorrow

- Acquiring Vocabulary
- Morphology
- Stemmers