POS tags 10/4/16

UMass CS 585

CS585 In-class exercise, 10/4/2016 Name: _____

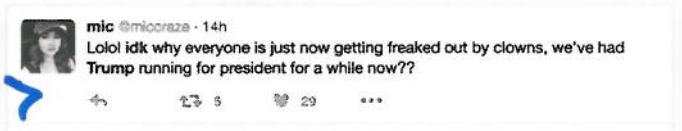
Tag these two sentences, using the PTB tagset below. Don't bother with punctuation, but tag all other tokens.

The New York Times

Based on Mr. Trump 's boasting and gaudy lifestyle

Mr. Wallach imagined he would soon be leading impressive

construction projects around the globe .



Lolol idk why everyone is just now getting freaked out by

clowns , we 've had Trump running for president for a

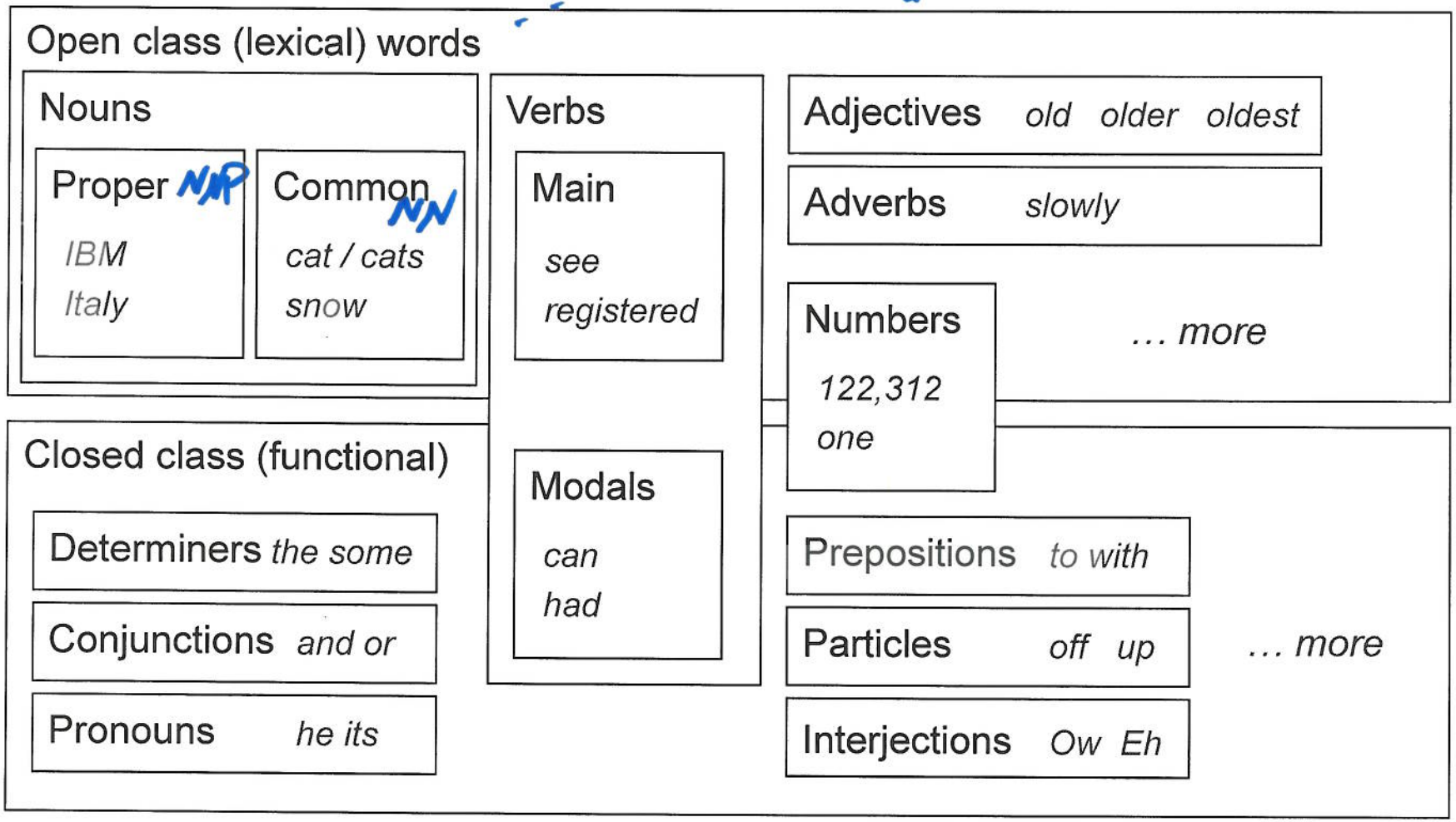
while now ??

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	*#*
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Figure 9.1 Penn Treebank part-of-speech tags (including punctuation).

Open vs closed classes

Content Words



slide credit: Chris Manning

Why POS?

- Good for downstream applications
  - Syntactic Parsing
  - Sentiment
  - Named Entity Recog.

POS patterns: simple noun phrases

- Quick and dirty noun phrase identification
  http://brenocon.com/JustesonKatz1995.pdf
  http://brenocon.com/handler2016phrases.pdf

Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A \mid N)^+ \mid ((A \mid N)^*(NP)^?)(A \mid N)^*)N$,

(Adj/* Noun+ (Prep Det? (Adj/Noun)+)

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

POS patterns: sentiment

- Turney (2002): identify bigram phrases, from unlabeled corpus, useful for sentiment analysis.

Table 1. Patterns of tags for extracting two-word phrases from reviews.

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table 2. An example of the processing of a review that the author has classified as *recommended*.[6]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732

(plus sentiment PMI stuff)

POS Tagging: lexical ambiguity

Can we just use a tag dictionary (one tag per word type)?

Types:	WSJ	Brown
Unambiguous (1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous (2+ tags)	7,025 (**14%**)	8,050 (**15%**)

Tokens:		
Unambiguous (1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous (2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Most words types are unambiguous ...

But not so for *tokens*!

- Ambiguous wordtypes tend to be very common ones.
  - I know **that** he is honest = IN (relativizer)
  - Yes, **that** play was nice = DT (determiner)
  - You can't go **that** far = RB (adverb)

How to build a POS tagger?

- Key sources of information:
  - 1. The word itself
  - 2. Word-internal characters
  - 3. POS tags of surrounding words: syntactic context