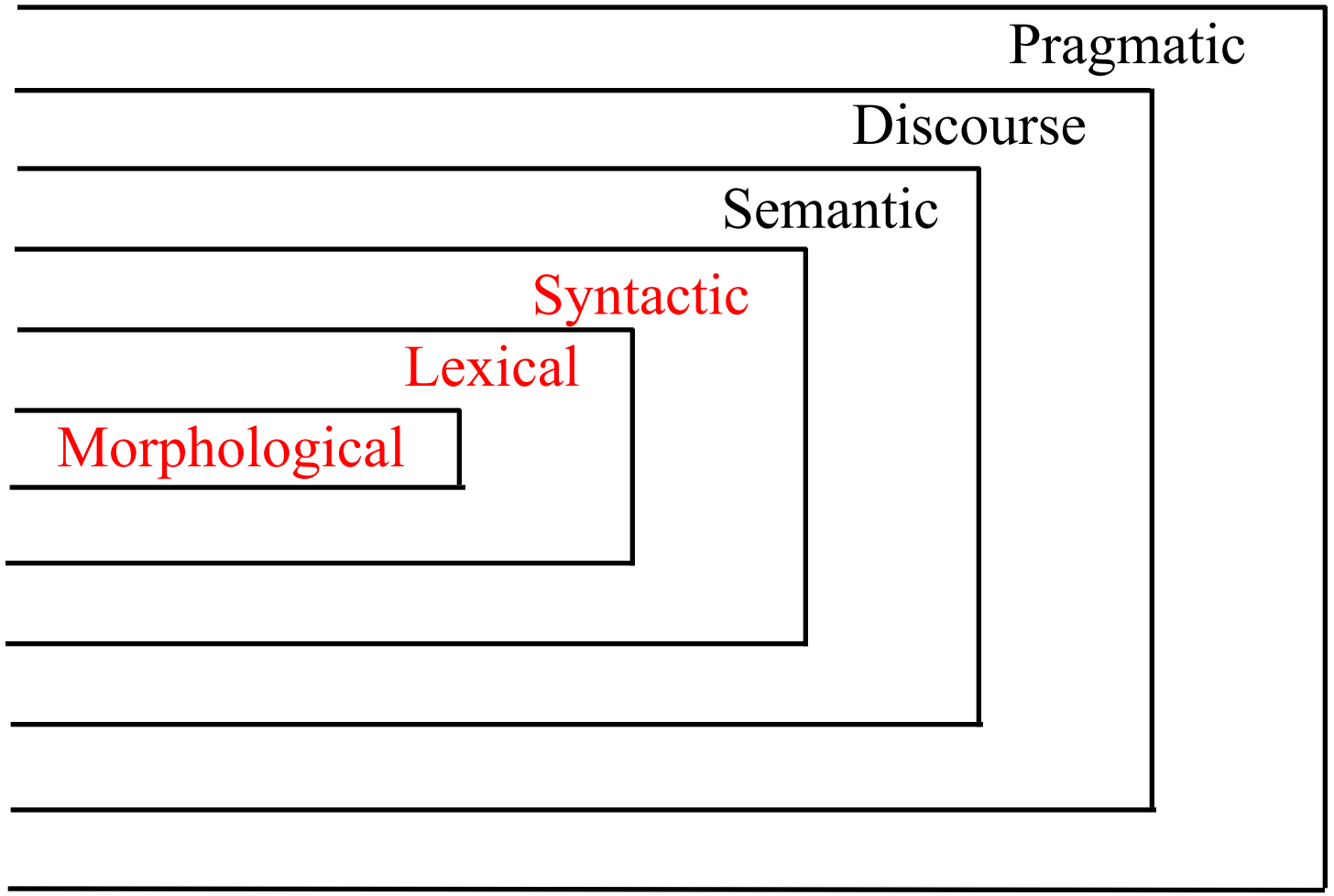
Introduction to Part-Of-Speech (POS) Tagging

Synchronic Model of Language

- POS tags are assigned to words, but may use adjacent words for information



What is Part-Of-Speech Tagging?

- The general purpose of a part-of-speech tagger is to associate each word in a text with its correct lexical-syntactic category (represented by a tag)

  03/14/1999 (AFP)… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

  … the|DT extremist|JJ Harkatul|NNP Jihad|NNP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NNP dissident|NN Osama|NNP bin|NN Laden|NNP …

What are Parts-of-Speech?

- Approximately 8 traditional basic word classes, sometimes called lexical classes or types
- These are the ones taught in grade school grammar
  - N noun *chair, bandwidth, pacing*
  - V verb *study, debate, munch*
  - ADJ adjective *purple, tall, ridiculous (includes articles)*
  - ADV adverb *unfortunately, slowly*
  - P preposition *of, by, to*
  - CON conjunction *and, but*
  - PRO pronoun *I, me, mine*
  - INT interjection *um*

Classes for Open Class Words

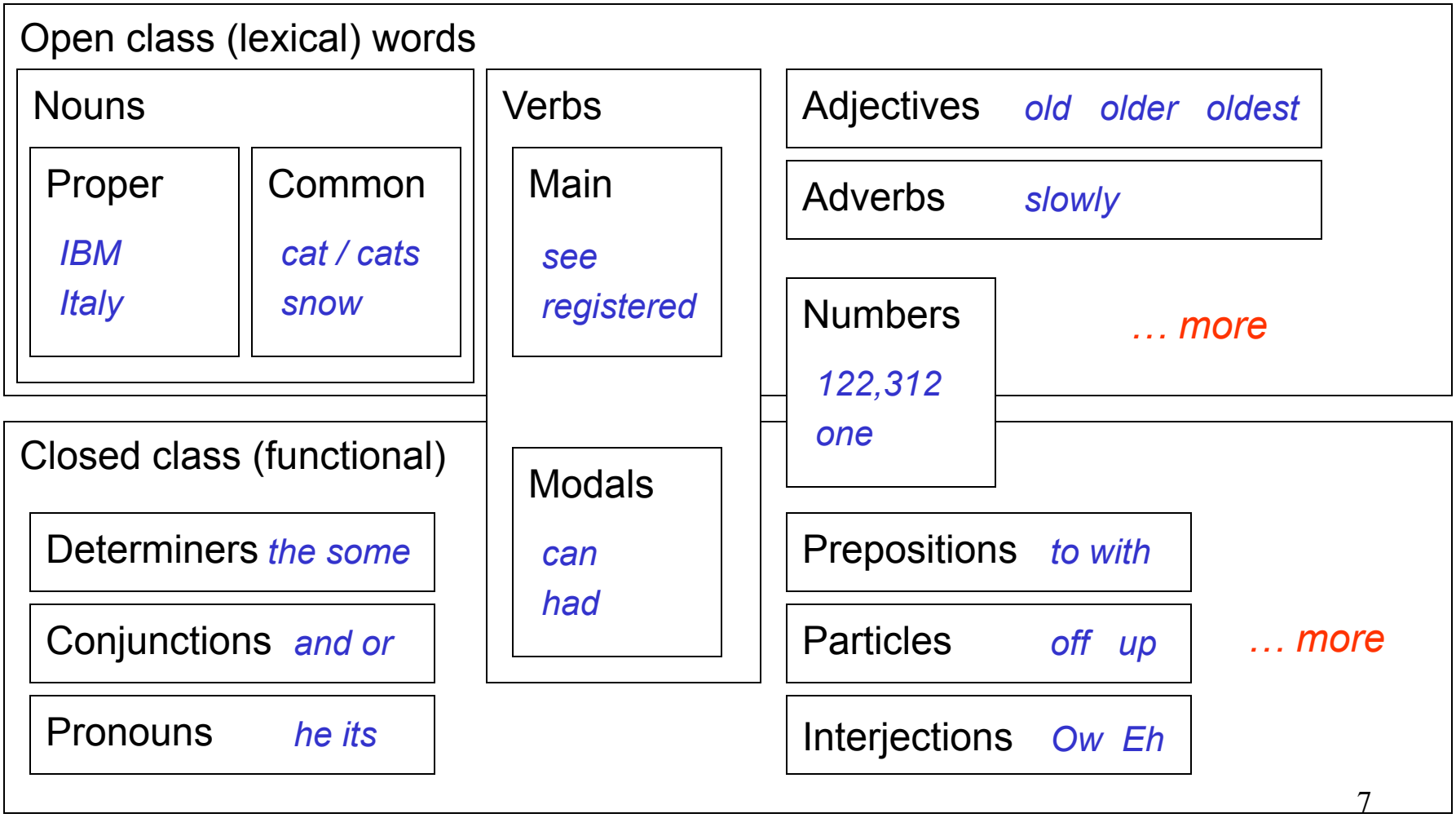
- Open classes – can add words to these basic word classes:
  - Nouns, Verbs, Adjectives, Adverbs.
- Nouns: people, places, things
  - Classes of nouns
    - proper (Boulder, Granby, Eli Manning) vs. common nouns
    - count (have plurals, get counted: chair/chairs, one chair, two chairs) vs. mass (don't get counted: furniture, snow, salt, communism)
  - Properties of nouns: can be preceded by a determiner, adjectives etc.
- Verbs: actions and processes
- Adjectives: properties, qualities
- Adverbs: hodgepodge! e.g. direction, degree, manner:
  - Unfortunately, John walked home extremely slowly yesterday
- Numerals, ordinals: one, two, three, third, …

Classes for Closed Class Words

- Closed classes– words are not added to these classes:
  - determiners: a, an, the
  - pronouns: she, he, I
  - prepositions: on, under, over, near, by, …
    - over the river and through the woods
  - particles: up, down, on, off, …
    - Used with verbs and have slightly different meaning than when used as a preposition
      - she turned the paper over
- Closed class words are often function words which have structuring uses in grammar:
    - of, it , and , you
- Differ more from language to language than open class words

Open and Closed Classes

- We may want to make more distinctions than 8 classes:



Prepositions from CELEX

- Prepositions show relationships between other words
- *Charts show words from the CELEX on-line dictionary with frequencies from the COBUILD corpus*

of	540,085	through	14,964	worth	1,563	pace	12
in	331,235	after	13,670	toward	1,390	nigh	9
for	142,421	between	13,275	plus	750	re	4
to	125,691	under	9,525	till	686	mid	3
with	124,965	per	6,515	amongst	525	o'er	2
on	109,129	among	5,090	via	351	but	0
at	100,169	within	5,030	amid	222	ere	0
by	77,794	towards	4,700	underneath	164	less	0
from	74,843	above	3,056	versus	113	midst	0
about	38,428	near	2,026	amidst	67	o'	0
than	20,210	off	1,695	sans	20	thru	0
over	18,071	past	1,575	circa	14	vice	0

Charts from Jurafsky and Martin text

English Single-Word Particles

- Definition of the term "particle" in linguistics varies
- Primarily words that used to provide shades of meaning to other words, particularly verbs

aboard	aside	besides	forward(s)	opposite	through
about	astray	between	home	out	throughout
above	away	beyond	in	outside	together
across	back	by	inside	over	under
ahead	before	close	instead	overhead	underneath
alongside	behind	down	near	past	up
apart	below	east, etc.	off	round	within
around	beneath	eastward(s),etc.	on	since	without

Pronouns in CELEX

- Personal
  - he, ours
- Demonstrative
  - that, those
- Reflexive
  - myself, ourselves
- Indefinite
  - one, neither, somebody, both

it	199,920	how	13,137	yourself	2,437	no one	106
I	198,139	another	12,551	why	2,220	wherein	58
he	158,366	where	11,857	little	2,089	double	39
you	128,688	same	11,841	none	1,992	thine	30
his	99,820	something	11,754	nobody	1,684	summat	22
they	88,416	each	11,320	further	1,666	suchlike	18
this	84,927	both	10,930	everybody	1,474	fewest	15
that	82,603	last	10,816	ourselves	1,428	thyself	14
she	73,966	every	9,788	mine	1,426	whomever	11
her	69,004	himself	9,113	somebody	1,322	whosoever	10
we	64,846	nothing	9,026	former	1,177	whomsoever	8
all	61,767	when	8,336	past	984	wherefore	6
which	61,399	one	7,423	plenty	940	whereat	5
their	51,922	much	7,237	either	848	whatsoever	4
what	50,116	anything	6,937	yours	826	whereon	2
my	46,791	next	6,047	neither	618	whoso	2
him	45,024	themselves	5,990	fewer	536	aught	1
me	43,071	most	5,115	hers	482	howsoever	1
who	42,881	itself	5,032	ours	458	thrice	1
them	42,099	myself	4,819	whoever	391	wheresoever	1
no	33,458	everything	4,662	least	386	you-all	1
some	32,863	several	4,306	twice	382	additional	0
other	29,391	less	4,278	theirs	303	anybody	0
your	28,923	herself	4,016	wherever	289	each other	0
its	27,783	whose	4,005	oneself	239	once	0
our	23,029	someone	3,755	thou	229	one another	0
these	22,697	certain	3,345	'un	227	overmuch	0
any	22,666	anyone	3,318	ye	192	such and such	0
more	21,873	whom	3,229	thy	191	whate'er	0
many	17,343	enough	3,197	whereby	176	whenever	0
such	16,880	half	3,065	thee	166	whereof	0
those	15,819	few	2,933	yourselves	148	whereto	0
own	15,741	everyone	2,812	latter	142	whereunto	0
us	15,724	whatever	2,571	whichever	121	whichsoever	0

Conjunctions

- Links words and phrases and gives relationship between them

and	514,946	yet	5,040	considering	174	forasmuch as	0
that	134,773	since	4,843	lest	131	however	0
but	96,889	where	3,952	albeit	104	immediately	0
or	76,563	nor	3,078	providing	96	in as far as	0
as	54,608	once	2,826	whereupon	85	in so far as	0
if	53,917	unless	2,205	seeing	63	inasmuch as	0
when	37,975	why	1,333	directly	26	insomuch as	0
because	23,626	now	1,290	ere	12	insomuch that	0
so	12,933	neither	1,120	notwithstanding	3	like	0
before	10,720	whenever	913	according as	0	neither nor	0
though	10,329	whereas	867	as if	0	now that	0
than	9,511	except	864	as long as	0	only	0
while	8,144	till	686	as though	0	provided that	0
after	7,042	provided	594	both and	0	providing that	0
whether	5,978	whilst	351	but that	0	seeing as	0
for	5,935	suppose	281	but then	0	seeing as how	0
although	5,424	cos	188	but then again	0	seeing that	0
until	5,072	supposing	185	either or	0	without	0

Auxiliary Verbs

- Auxiliary, or helping verbs, are used with main verbs to express time or mood
  - Modal verbs are the auxiliary verbs that express likelihood or ability
    - Can, might, must, could, should, …

can	70,930	might	5,580	shouldn't	858
will	69,206	couldn't	4,265	mustn't	332
may	25,802	shall	4,118	'll	175
would	18,448	wouldn't	3,548	needn't	148
should	17,760	won't	3,100	mightn't	68
must	16,520	'd	2,299	oughtn't	44
need	9,955	ought	1,845	mayn't	3
can't	6,375	will	862	dare	??
have	???				

Possible Tag Sets for English

- Kucera & Francis (Brown Corpus) – 87 POS tags
- C5 (British National Corpus) – 61 POS tags
  - Tagged by Lancaster's UCREL project
- Penn Treebank – 45 POS tags
  - Most widely used of the tag sets today

Penn Treebank

- A corpus containing:
  - over 1.6 million words of hand-parsed material from the Dow Jones News Service, plus an additional 1 million words tagged for part-of-speech.
  - the first fully parsed version of the Brown Corpus, which has also been completely retagged using the Penn Treebank tag set.
  - source code for several software packages which permits the user to search for specific constituents in tree structures.
- Costs $1,250 to $2,500 for research use
- Separate licensing needed for commercial use

Word Classes: Penn Treebank Tag Set

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	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, singular	*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*			

Examples of Penn Treebank Tagging

- The/DT grand/JJ jury/NN commented/VBD
  on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Book/VB that/DT flight/NN ./.
- Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/?