



# Applied Natural Language Processing

Info 256

Lecture 19: POS Tagging (Oct. 30, 2023)

David Bamman, UC Berkeley

context

everyone likes

\_\_\_\_\_

a bottle of

\_\_\_\_\_

is on the table

\_\_\_\_\_ makes you drunk

a cocktail with

\_\_\_\_\_

and seltzer

# Distribution

- Words that appear in similar contexts have similar representations (and similar **meanings**, by the distributional hypothesis).

# Parts of speech

- Parts of speech are categories of words defined **distributionally** by the morphological and syntactic contexts a word appears in.

# Morphological distribution

POS often defined by distributional properties; verbs = the class of words that each combine with the same set of affixes

	-s	-ed	-ing
walk	walks	walked	walking
slice	slices	sliced	slicing
believe	believes	believed	believing
of	*ofs	*ofed	*ofing
red	*reds	*redded	*reding

# Morphological distribution

We can look to the function of the affix (denoting past tense) to include irregular inflections.

	-s	-ed	-ing
walk	walks	walked	walking
sleep	sleeps	slept	sleeping
eat	eats	ate	eating
give	gives	gave	giving

# Syntactic distribution

- Substitution test: if a word is replaced by another word, does the sentence remain **grammatical**?

Kim saw the	elephant	before we did
	dog	
	idea	
	*of	
	*goes	

# Syntactic distribution

- These can often be too strict; some contexts admit substitutability for some pairs but not others.

Kim saw the elephant before we did

\*Sandy

both nouns but common  
vs. proper

both verbs but transitive  
vs. intransitive

Kim \*arrived the elephant before we did



Nouns	People, places, things, actions-made-nouns (“I like <b>swimming</b> ”). Inflected for singular/plural
Verbs	Actions, processes. Inflected for tense, aspect, number, person
Adjectives	Properties, qualities. Usually modify nouns
Adverbs	Qualify the manner of verbs (“She ran <b>downhill extremely quickly yesterday</b> ”)
Determiner	Mark the beginning of a noun phrase (“ <b>a</b> dog”)
Pronouns	Refer to a noun phrase (he, she, it)
Prepositions	Indicate spatial/temporal relationships ( <b>on</b> the table)
Conjunctions	Conjoin two phrases, clauses, sentences (and, or)

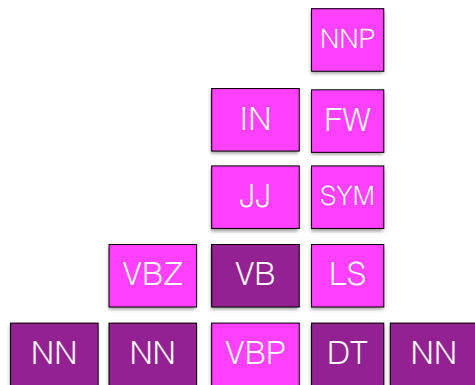
Open class

Nouns	fax, affluenza, subtweet, bitcoin, cronut, emoji, listicle, mocktail, selfie, skort
Verbs	text, chillax, manspreading, photobomb, unfollow, google
Adjectives	crunk, amazeballs, post-truth, woke
Adverbs	hella, wicked
Determiner	OOV? Guess Noun
Pronouns	
Prepositions	English has a new preposition, because internet [Garber 2013; Pullum 2014]
Conjunctions	

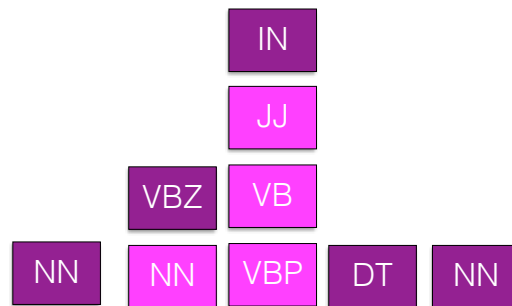
Closed class

# POS tagging

Labeling the tag that's correct for the context.



Fruit flies like a banana



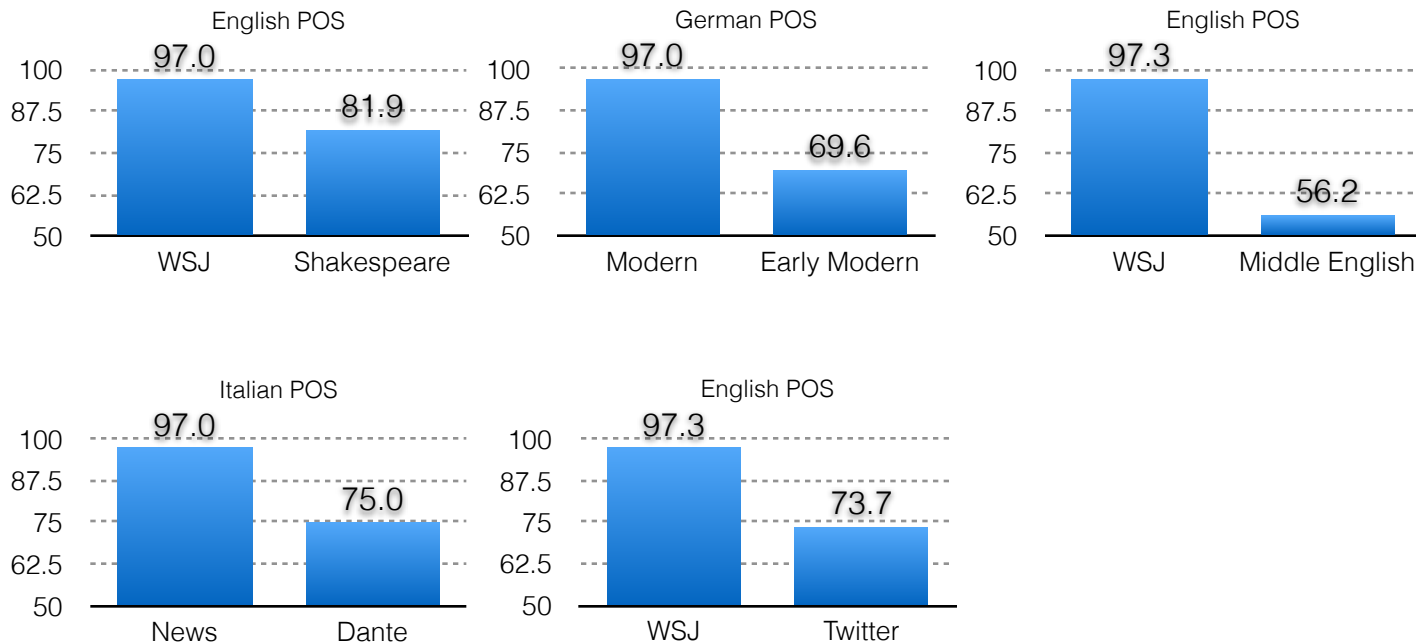
Time flies like an arrow

(Just tags in evidence within the Penn Treebank — more are possible!)

# State of the art

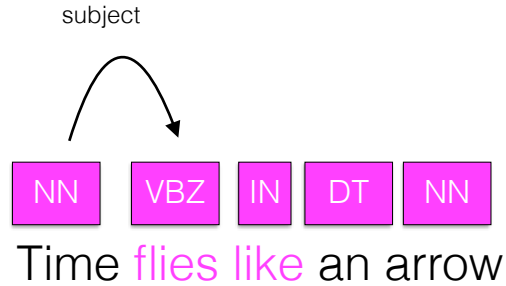
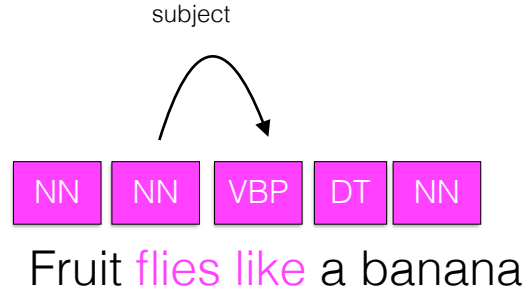
- Baseline: Most frequent class = 92.34%
- Token accuracy: 97% (English news)  
[Toutanova et al. 2003; Søgaard 2010]
  - Optimistic: includes punctuation, words with only one tag (deterministic tagging)
  - Substantial drop across domains (e.g., train on news, test on literature)
- Whole sentence accuracy: 55%

# Domain difference



Why is part of speech tagging useful?

# POS indicative of syntax



# POS indicative of MWE

at least one adjective/noun or noun phrase

and definitely  
one noun

$$((A | N)^+ | ((A | N)^*(NP))(A | N)^*)N$$

- AN*: linear function; lexical ambiguity; mobile phase  
*NN*: regression coefficients; word sense; surface area  
*AAN*: Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase  
*ANN*: cumulative distribution function; lexical ambiguity resolution; accessible surface area  
*NAN*: mean squared error; domain independent set; silica based packing  
*NNN*: class probability function; text analysis system; gradient elution chromatography  
*NPN*: degrees of freedom; [*no example*]; energy of adsorption



# POS is indicative of pronunciation

Noun	Verb
My conduct is great	I conduct myself well
She won the contest	I contest the ticket
He is my escort	He escorted me
That is an insult	Don't insult me
Rebel without a cause	He likes to rebel
He is a suspect	I suspect him

# Filtering

- Keyphrase extraction: select the top K terms that best describe a document; often only want **nouns**.

# Tagsets

- Penn Treebank
- Universal Dependencies
- Twitter POS

# Verbs

tag	description	example
VB	base form	I want to <b>like</b>
VBD	past tense	I/we/he/she/you <b>liked</b>
VBG	present participle	He was <b>liking</b> it
VBN	past participle	I had <b>liked</b> it
VBP	present (non 3rd-sing)	I <b>like</b> it
VBZ	present (3rd-sing)	He <b>likes</b> it
MD	modal verbs	He <b>can</b> go

# Nouns

non-proper

proper

tag	description	example
NN	non-proper, singular or mass	the company
NNS	non-proper, plural	the companies
NNP	proper, singular	Carolina
NNPS	proper, plural	Carolinas

# DT (Article)

- Articles (a, the, every, no)
- Indefinite determiners  
(another, any, some, each)
- That, these, this, those when preceding noun
- All, both when not preceding another determiner or possessive pronoun

65548	the/dt
26970	a/dt
4405	an/dt
3115	this/dt
2117	some/dt
2102	that/dt
1274	all/dt
1085	any/dt
953	no/dt
778	those/dt

# JJ (Adjectives)

- General adjectives

- *happy person*
- *new mail*

- Ordinal numbers

- *fourth person*

2002 other/jj  
1925 new/jj  
1563 last/jj  
1174 many/jj  
1142 such/jj  
1058 first/jj  
824 major/jj  
715 federal/jj  
698 next/jj  
644 financial/jj

# RB (Adverb)

- Most words that end in **-ly**
- Degree words (**quite, too, very**)
- Negative markers: **not, n't, never**

4410 n't/rb  
2071 also/rb  
1858 not/rb  
1109 now/rb  
1070 only/rb  
1027 as/rb  
961 even/rb  
839 so/rb  
810 about/rb  
804 still/rb



# IN (preposition, subordinating conjunction)

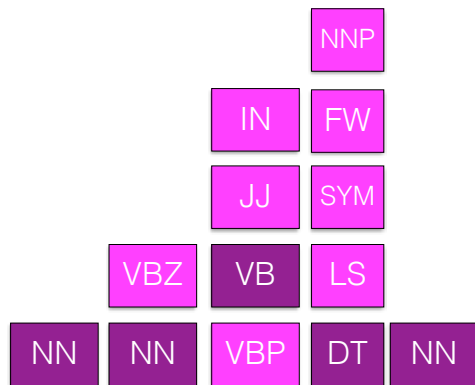
- All prepositions (except *to*) and subordinating conjunctions

- He jumped **on** the table  
**because** he was excited

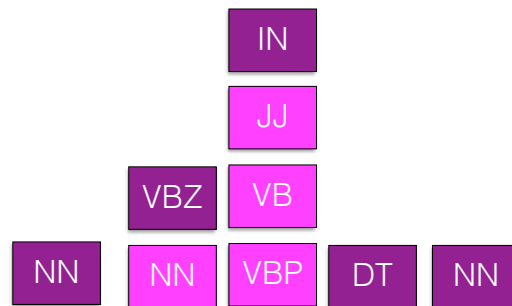
31111 of/in  
22967 in/in  
11425 for/in  
7181 on/in  
6684 that/in  
6399 at/in  
6229 by/in  
5940 from/in  
5874 with/in  
5239 as/in

# POS tagging

Labeling the tag that's correct for the context.



Fruit flies like a banana



Time flies like an arrow

(Just tags in evidence within the Penn Treebank — more are possible!)

# Sequence labeling

$$x = \{x_1, \dots, x_n\}$$

$$y = \{y_1, \dots, y_n\}$$

- For a set of inputs  $x$  with  $n$  sequential time steps, one corresponding label  $y_i$  for each  $x_i$

# Activity

- [11.pos/POS tagging.ipynb](#)