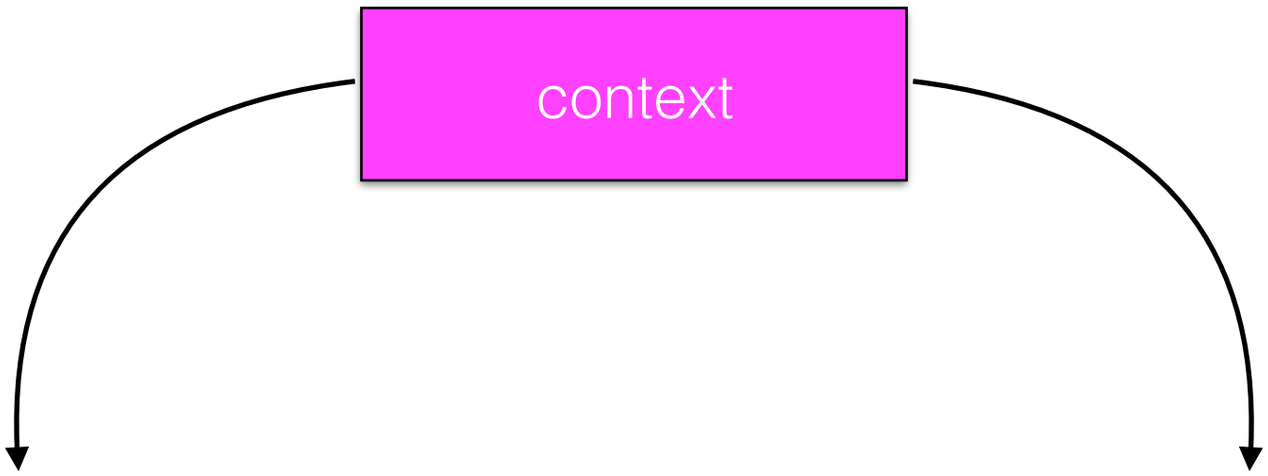
Natural Language Processing

Mehmet Can Yavuz, PhD
Adapted from Info 256 - David Bamman, UC Berkeley



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

Bender 2013

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

Bender 2013

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	

Bender 2013

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	
Kim *arrived the	elephant	before we did

both nouns but common vs. proper

both verbs but transitive vs. intransitive

Bender 2013

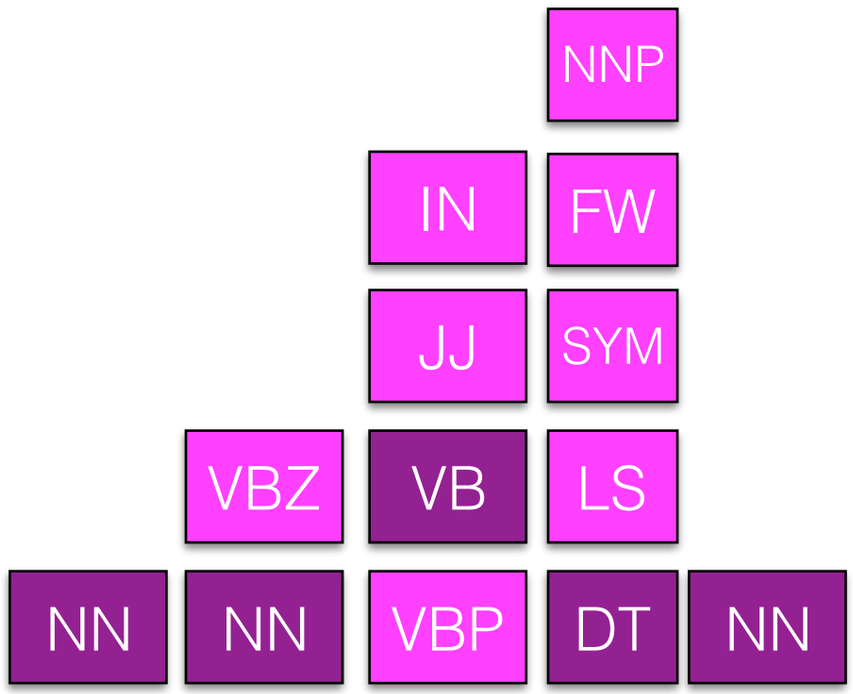
Nouns	People, places, things, actions-made-nouns (“I like swimming”). Inflected for singular/plural
Verbs	Actions, processes. Inflected for tense, aspect, number, person
Adjectives	Properties, qualities. Usually modify nouns
Adverbs	Qualify the manner of verbs (“She ran downhill extremely quickly yesteray”)
Determiner	Mark the beginning of a noun phrase (“a dog”)
Pronouns	Refer to a noun phrase (he, she, it)
Prepositions	Indicate spatial/temporal relationships (on 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
Closed class	Determiner	
	Pronouns	
	Prepositions	English has a new preposition, because internet [Garber 2013; Pullum 2014]
	Conjunctions	

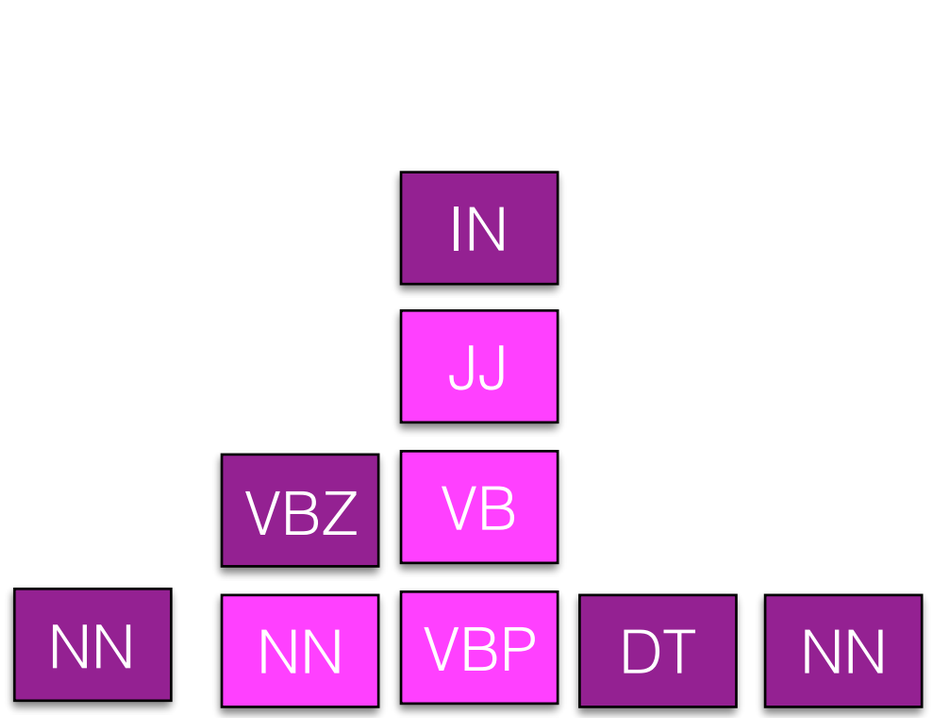
OOV? Guess Noun

POS tagging

Labeling the tag that's correct for the context.



Fruit flies like a banana



Time flies like an arrow

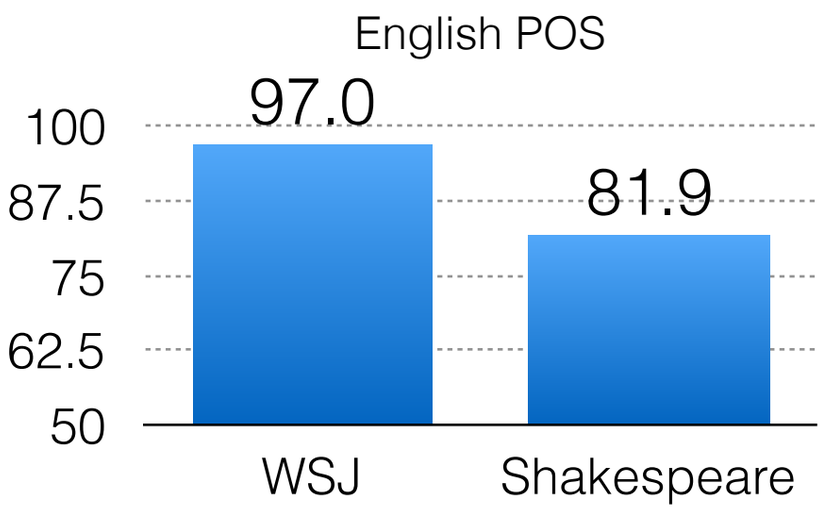
(Just tags in evidence within the Penn Treebank — more are possible!)

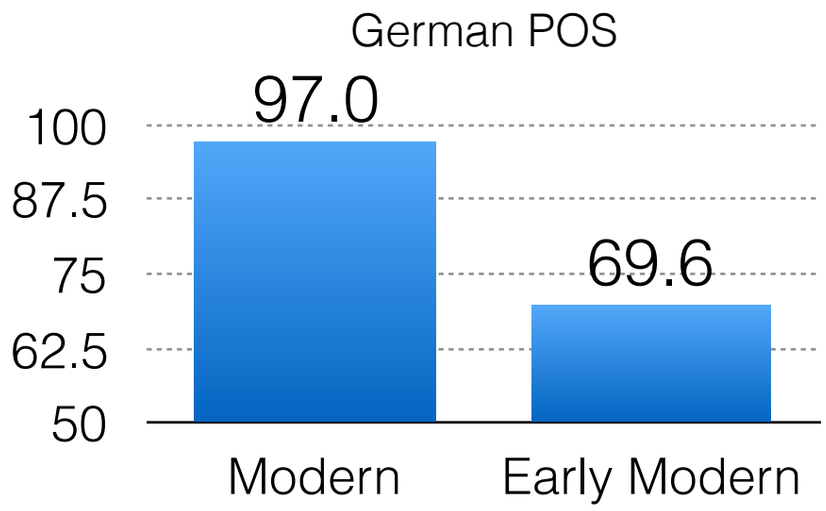
State of the art

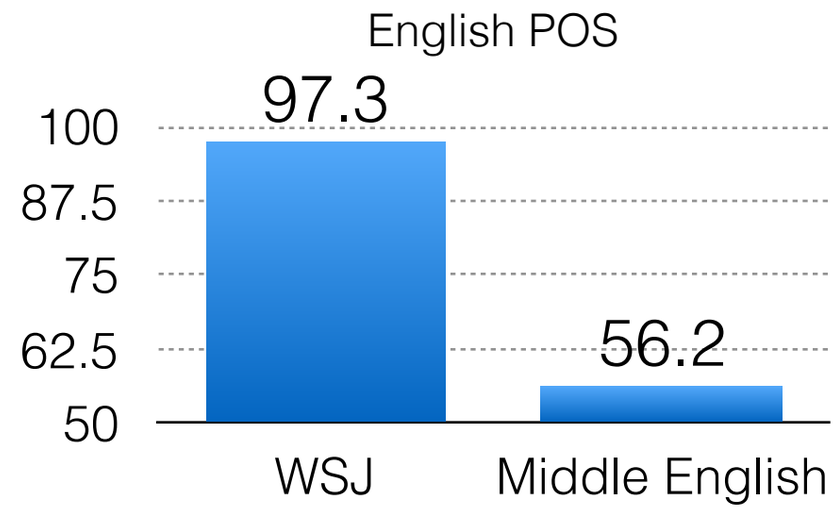
- Baseline: Most frequent class = 92.34%
- Token accuracy: 97% (English news)
  [Toutanova et al. 2003; Søgaard 2010]
  - Optimistic: includes punctuation, words with only one tag (deterministic tagging)
  - Substantial drop across domains (e.g., train on news, test on literature)
- Whole sentence accuracy: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

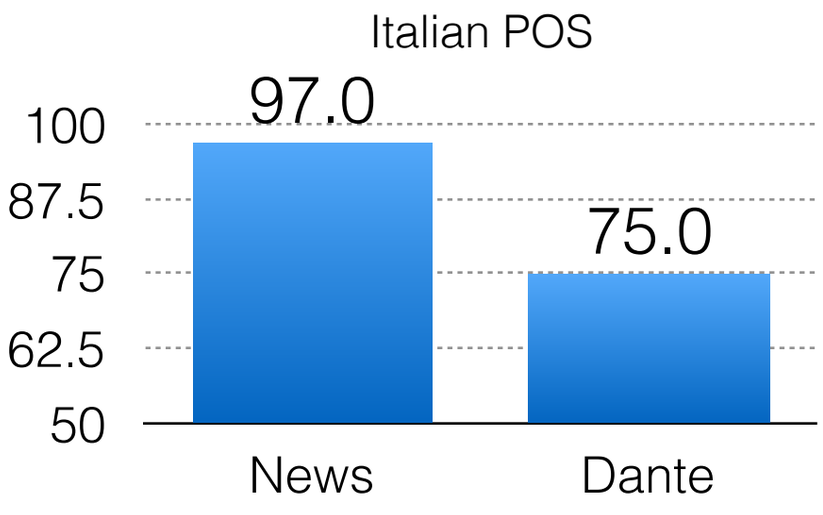
Manning 2011

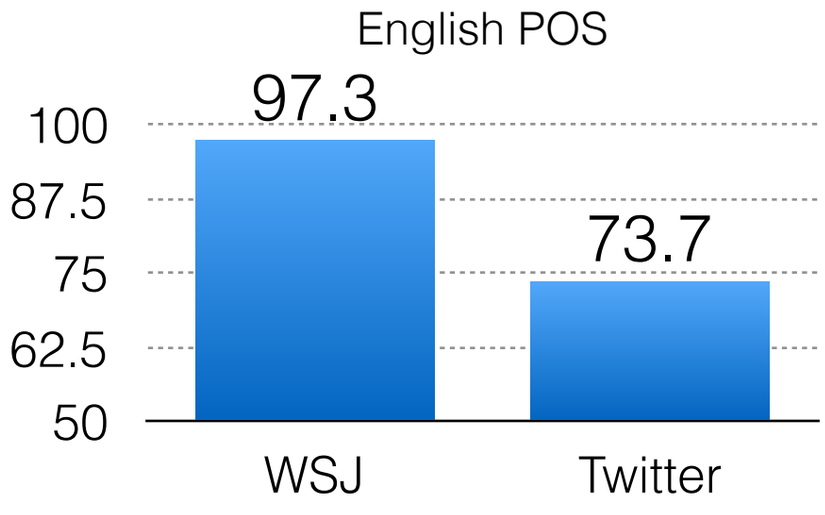
Domain difference











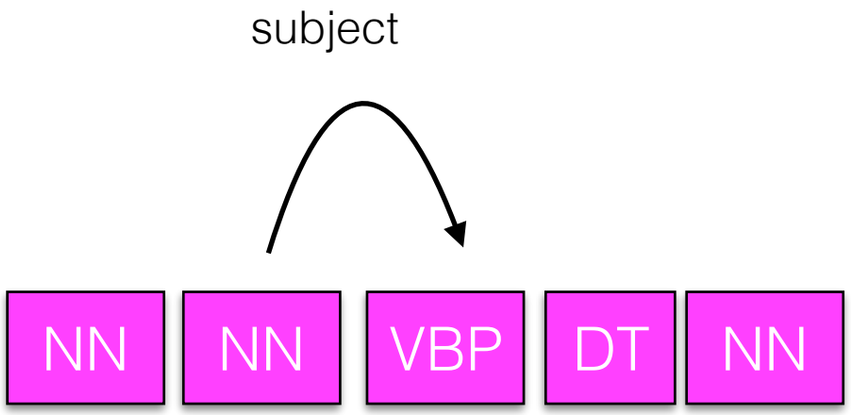
Sources of error

Lexicon gap	4.5%	a 60% slash/NN the common stock dividend
Unknown word	4.5%	blaming the disaster on substandard/JJ construction
Could plausibly get right	16.0%	market players overnight/RB in Tokyo began bidding up oil prices
Difficult linguistics	19.5%	They set/VBP up absurd situations, detached from reality
Underspecified/unclear	12.0%	a $ 10 million fourth-quarter charge against/IN discontinued/JJ operations
Inconsistent/no standard	28.0%	Orson Welles 's Mercury Theater in the '30s/NNS
Gold standard wrong	15.5%	Our market got hit/VB a lot harder on Monday than the listed market

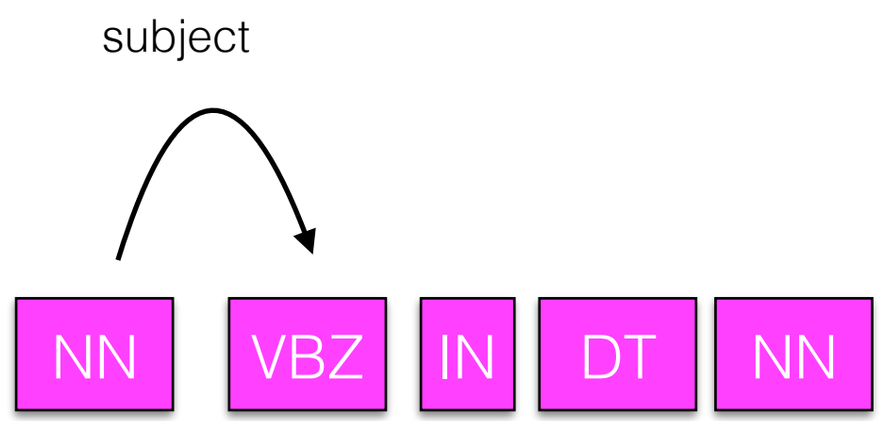
Manning 2011

Why is part of speech tagging useful?

POS indicative of syntax



Fruit flies like a banana



Time flies like an arrow

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

Justeson and Katz 1995

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 like
VBD	past tense	I/we/he/she/you liked
VBG	present participle	He was liking it
VBN	past participle	I had liked it
VBP	present (non 3rd-sing)	I like it
VBZ	present (3rd-sing)	He likes it
MD	modal verbs	He can 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

Santorini 1990

```
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
```

Santorini 1990

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
```

Santorini 1990

IN (preposition, subordinating conjunction)

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

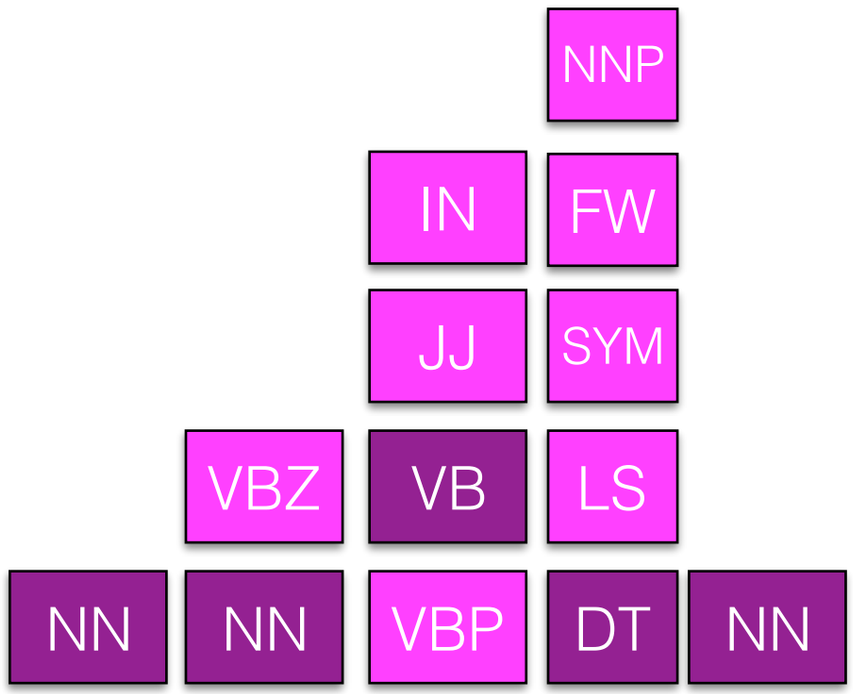
- He jumped on the table because he was excited

```
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
```

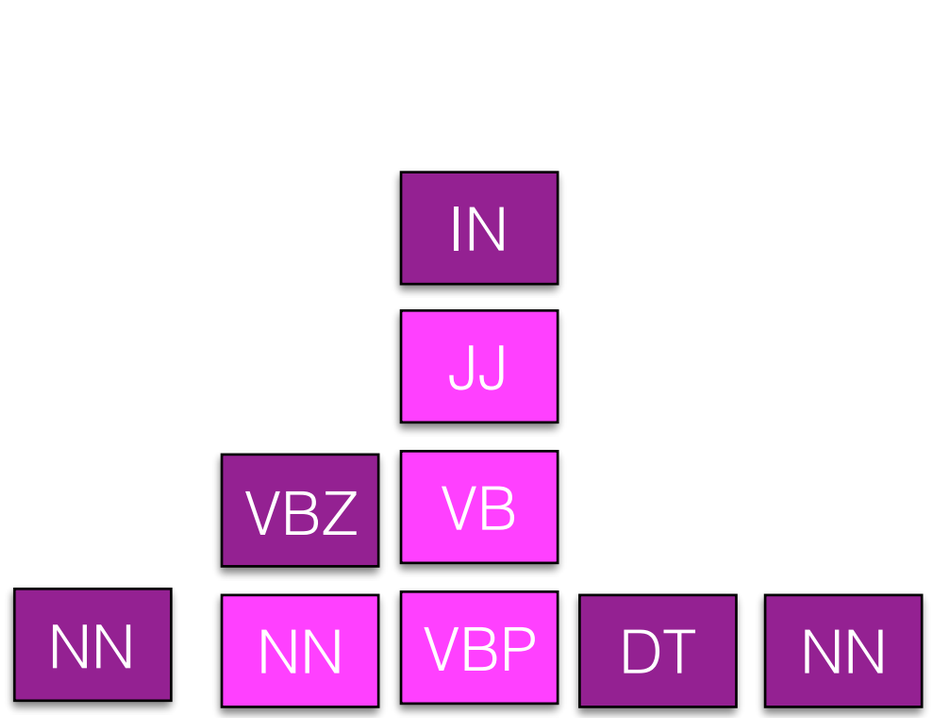
Santorini 1990

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

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

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

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