Outline of POS/HMM unit:

- The POS-tagging problem
- Hidden Markov Models definition
- HMM Problem 1 and the forward algorithm
- HMM Problem 2 and the Viterbi algorithm, applied to POS-tagging
- HMM Problem 3 and the Baum-Welch algorithm, with other linguistic applications

English parts of speech:

Noun	Adjective	Pronoun	Conjunction
Verb	Adverb	Preposition	Interjection

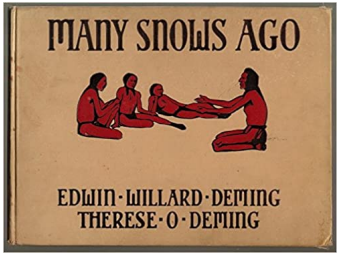
Ancient Indian (Sanskrit) parts of speech:

Noun	Verb	Preverb	Particle

Ancient Greek parts of speech:

Noun	Participle	Pronoun	Conjunction
Verb	Adverb	Preposition	Article

Many languages, including English, divide common nouns into count nouns and mass nouns. Count nouns can occur in the singular and plural and can be counted. Mass nouns are used when something is conceptualized as a homogenous group. So *snow*, *salt*, and *communism* are not counted (i.e., **two snows* or **two communisms*). J&M p 149



Karl Marx and Josef Stalin represent two very different communisms.

Nouns and adjectives

Nouns can be used attributively:

> *There is of course nothing new in putting a noun to this use when no convenient adjective is available; examples abound in everyday speech—***government department**, **nursery school**, **television set**, **test match**, *and innumerable others. But the noun-adjective, useful in its proper place, is now running riot and corrupting the language.* *H.W. Fowler, Modern English Usage*

Adjectives can be used substantively:

> *Do not let the* **perfect** *be the enemy of the* **good**.

Word classes

Linguists group the words of a language into classes (sets) which show similar syntactic behavior, and often a typical semantic type. These word classes are otherwise called **syntactic** *or* **grammatical categories**, *but more commonly still by the traditional name* **parts of speech** *(POS). Three important parts of speech are* **noun**, **verb**, *and* **adjective**. *. . . The most basic test for words belonging to the same class is the* **substitution test**. *Adjectives can be picked out as words that occur in the frame:*

$$\textit{The} \left\{ \begin{array}{c} \textit{sad} \\ \textit{intelligent} \\ \textit{green} \\ \textit{fat} \\ \ldots \end{array} \right\} \textit{one is in the corner}$$

Manning and Schütze, Foundations of Statistical NLP, pg 81

Computed word classes

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
June March July April January December October November September August
people guys folks fellows CEOs chaps doubters commies unfortunates blokes
down backwards ashore sideways southward northward overboard aloft downwards adrift
water gas coal liquid acid sand carbon steam shale iron
great big vast sudden mere sheer gigantic lifelong scant colossal
man woman boy girl lawyer doctor guy farmer teacher citizen
American Indian European Japanese German African Catholic Israeli Italian Arab
pressure temperature permeability density porosity stress velocity viscosity gravity tension
mother wife father son husband brother daughter sister boss uncle
machine device controller processor CPU printer spindle subsystem compiler plotter
John George James Bob Robert Paul William Jim David Mike
anyone someone anybody somebody
feet miles pounds degrees inches barrels tons acres meters bytes
director chief professor commissioner commander treasurer founder superintendent dean custodian
liberal conservative parliamentary royal progressive Tory provisional separatist federalist PQ
had hadn't hath would've could've should've must've might've
asking telling wondering instructing informing kidding reminding bothering thanking deposing
that tha theat
head body hands eyes voice arm seat eye hair mouth

Table 2
Classes from a 260,741-word vocabulary.

Brown et al, "Class-Based *n*-gram Models of Natural Language"

Computed word classes

little prima moment's trifle tad Litle minute's tinker's hornet's teammate's
6
ask remind instruct urge interrupt invite congratulate commend warn applaud
object apologize apologise avow whish
cost expense risk profitability deferral earmarks capstone cardinality mintage reseller
B dept. AA Whitey CL pi Namerow PA Mgr. LaRose
Rel rel. #S Shree
S Gens nai Matsuzawa ow Kageyama Nishida Sumit Zollner Mallik
research training education science advertising arts medicine machinery Art AIDS
rise focus depend rely concentrate dwell capitalize embark intrude typewriting
Minister mover Sydneys Minster Miniter
3
running moving playing setting holding carrying passing cutting driving fighting
court judge jury slam Edelstein magistrate marshal Abella Scalia larceny
annual regular monthly daily weekly quarterly periodic Good yearly convertible
aware unaware unsure cognizant apprised mindful partakers
force ethic stoppage force's conditioner stoppages conditioners waybill forwarder Atonabee
systems magnetics loggers products' coupler Econ databanks Centre inscriber correctors
industry producers makers fishery Arabia growers addiction medalist inhalation addict
brought moved opened picked caught tied gathered cleared hung lifted

Table 3
Randomly selected word classes.

Brown et al, "Class-Based *n*-gram Models of Natural Language"

Universal		Penn Treebank		
ADJ	Adjective	JJ	Adjective	*yellow*
		JJR	Comparative adjective	*bigger*
		JJS	Superlative adjective	*wildest*
ADP	Adposition	IN	Preposition	*of, in , by*
		RP	Particle	*up, off*
ADV	Adverb	RB	Adverb	*quickly*
		RBR	Comparative adverb	*faster*
		RBS	Superlative adverb	*fastest*
		WRB	Wh-adverb	*how, where*
CONJ	Conjunction	CC	Coordinating conjunction	*and, but, or*

	Universal		**Penn Treebank**	
DET	Determiner, article	DT	Determiner	*a, the*
		PDT	Predeterminer	*all, both*
		PRP$	Posessive pronoun	*your, one's*
		WDT	Wh-determiner	*which, that*
		WP$	Wh-possessive	*whose*
NOUN	Noun	NN	Singular or mass noun	*llama*
		NNP	Proper noun, singular	*IBM*
		NNPS	Noun, plural	*llamas*
NUM	Numeral	CD	Cardinal number	*one, two*
PRT	Particle	POS	Possessive ending	*'s*
		TO	"to" [Infinitive marker]	*to*
PRON	Pronoun	EX	Existential "there"	*there*
		PRP	Personal pronoun	*I, you, he*
		WP	Wh-pronoun	*what, who*

	Universal		**Penn Treebank**	
VERB	Verb	MD	Modal *can, should*	
		VB	Verb base	*eat*
		VBD	Verb past tense	*ate*
		VBG	Verb gerund	*eating*
		VBN	Verb past participle	*eaten*
		VBP	Verb non-3sp	*eat*
		VBZ	Verb 3sp	*eats*
.	Puntuation mark	(none)		
X	Other	FW	Foreign word	*mea culpa*
		LS	List item marker	*1, 2, One*
		SYM	Symbol	*+, %, &*
		UH	Interjection	*ah, oops*