

Lexical classes and HMMs

CSCI 384: Computational Linguistics

CSCI 384

October 20, 2017

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:*

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

Manning and Schütze, Foundations of Statistical NLP, pg 81

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**.*

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 cus-
todian
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 Little minute's tinker's hornet's teammate's

6

ask remind instruct urge interrupt invite congratulate commend warn applaud

object apologize apologise avow wish

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 Minitier

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

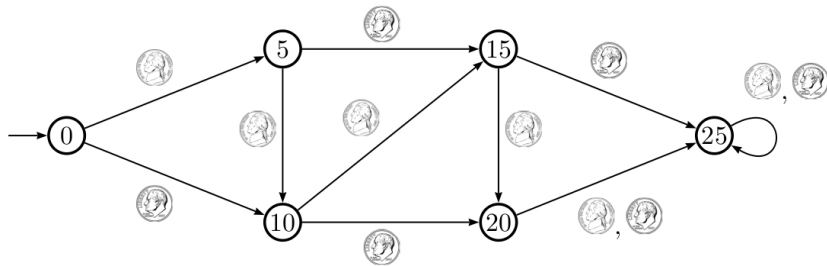
brought moved opened picked caught tied gathered cleared hung lifted

Table 3

Randomly selected word classes.

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

Finite state machines



Finite state machines

