PART-OF-SPEECH (POS)

WOA7013

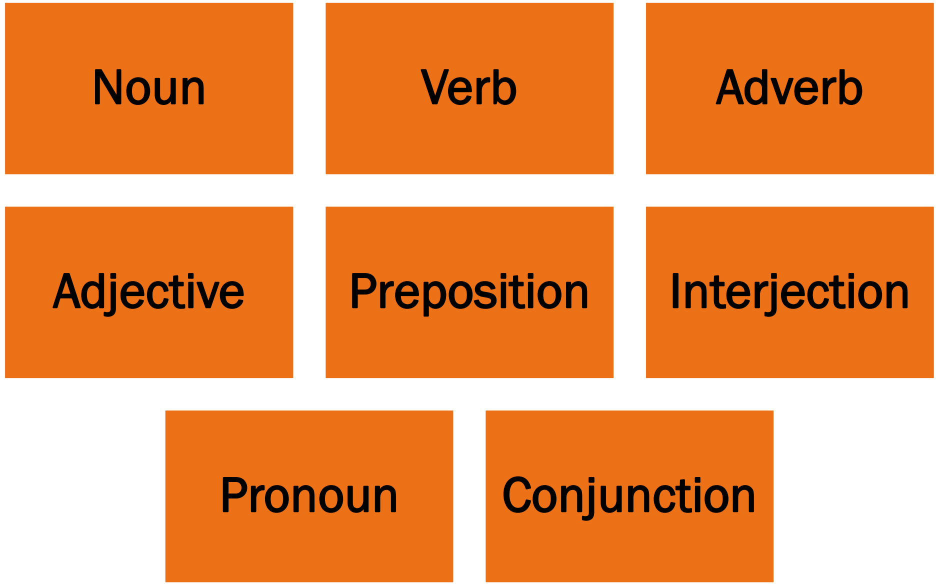
THEORY AND APPLICATIONS OF NATURAL LANGUAGE PROCESSING

Introduction

Starting with Aristotle in the West, there was the idea of having parts-of-speech a.k.a,

- lexical categories, word classes, tags, POS

Basically, there are 8 parts of speech:



Noun

Verb

Adverb

Adjective

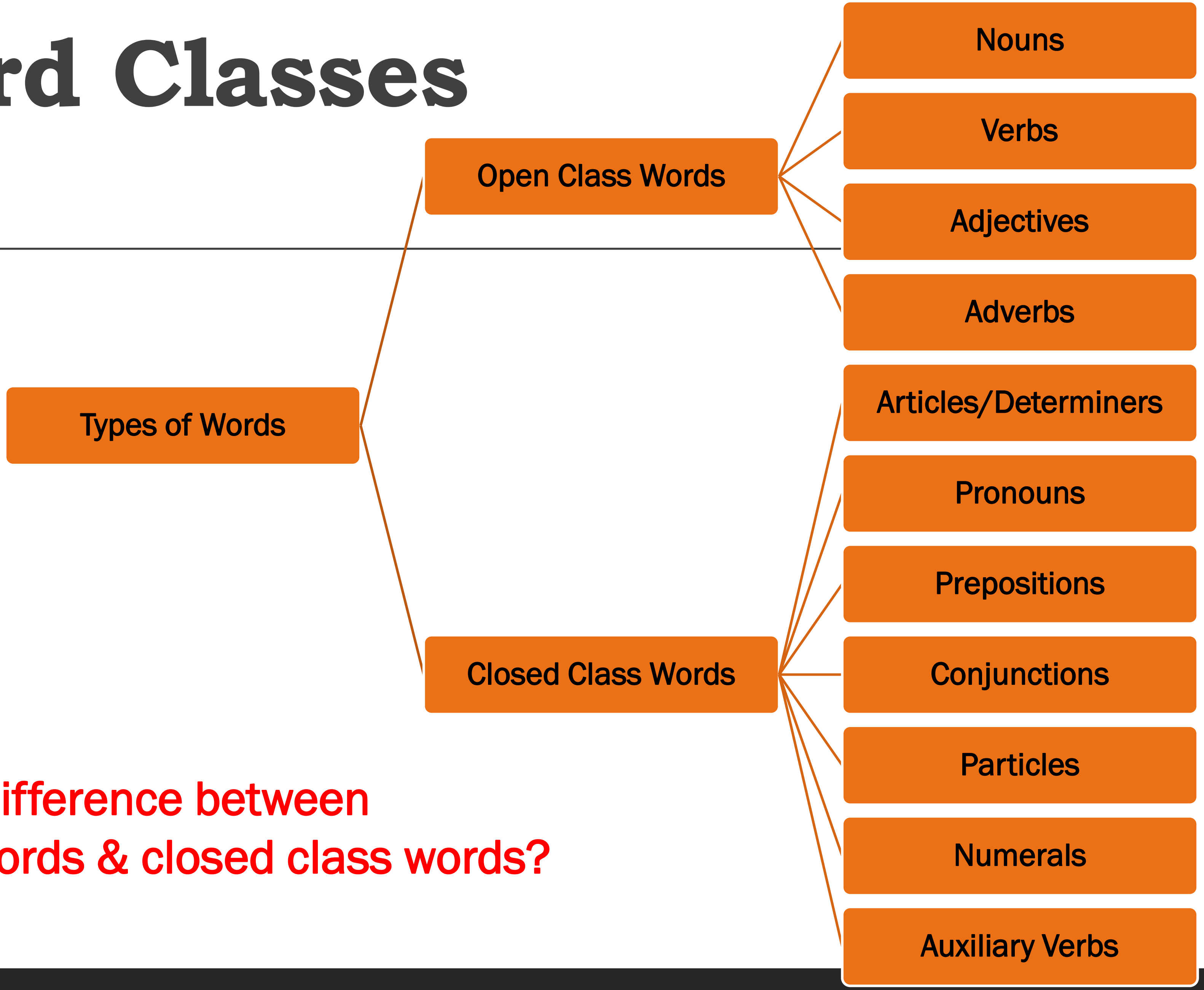
Preposition

Interjection

Pronoun

Conjunction

English Word Classes



- Types of Words
  - Open Class Words
    - Nouns
    - Verbs
    - Adjectives
    - Adverbs
  - Closed Class Words
    - Articles/Determiners
    - Pronouns
    - Prepositions
    - Conjunctions
    - Particles
    - Numerals
    - Auxiliary Verbs

What is the difference between
open class words & closed class words?

- Nouns
  - Proper Nouns
    - A word or group of words that is the name of a person or a place.
    - E.g. *IBM, Malaysia*
  - Common Nouns
    - A word or group of words that is a thing or activity, or a quality or idea.
    - E.g. *school, cat, football*
  - Can be used as the subject or object of a verb.

▣ Verb

- ❖ A word or group of words that is used in describing an action, experience, or state.
- ❖ Can be divided into several different classes:
  - ❖ Main verbs
  - ❖ Auxiliary verbs *(closed class words)*
- ❖ E.g:
  - ❖ *Run, fly, walk, shock, feel*.

Adjective

- A word or group of words that describes a noun or pronoun.
- E.g:
  - *Crazy, black, amaze, smart*, etc.

Adverb

- A word or group of words that describes or adds to the meaning of a verb.
- E.g:
  - *Slowly, hungrily, away, naturally*, etc.

Match the Word Classes and their examples

- Determiners
- Pronouns
- Prepositions
- Conjunctions
- Particles
- Numerals

A. up, down, on, off, in, out

B. a, an, the

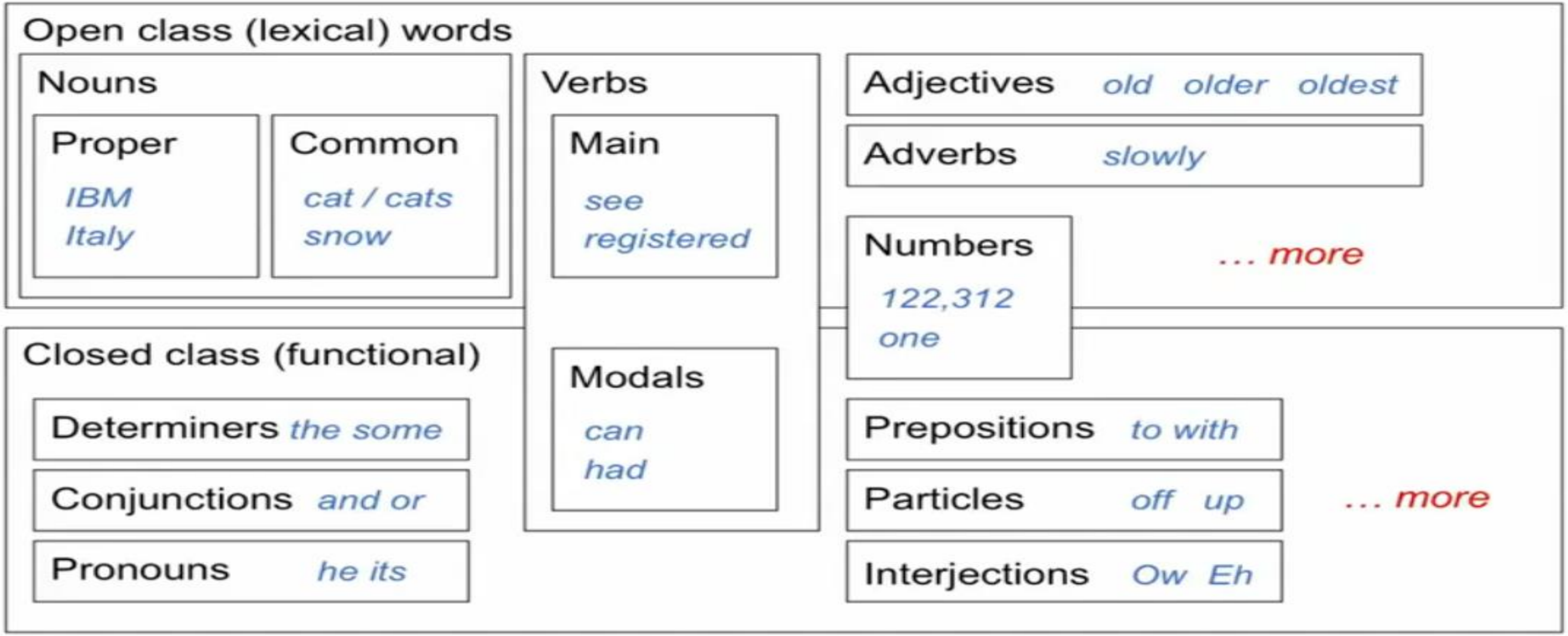
C. she, I, we

D. on, under, over, near, by, at, from

E. one, two, first, second

F. and, but, or, as, since, because

SUMMARY – PARTS-OF-SPEECH



Open class (lexical) words

- Nouns
  - Proper: *IBM*, *Italy*
  - Common: *cat / cats*, *snow*
- Verbs
  - Main: *see*, *registered*
  - Modals: *can*, *had*
- Adjectives: *old older oldest*
- Adverbs: *slowly*
- Numbers: *122,312*, *one*
- *... more*

Closed class (functional)

- Determiners: *the some*
- Conjunctions: *and or*
- Pronouns: *he its*
- Prepositions: *to with*
- Particles: *off up*
- Interjections: *Ow Eh*
- *... more*

The Penn Treebank POS Tagset

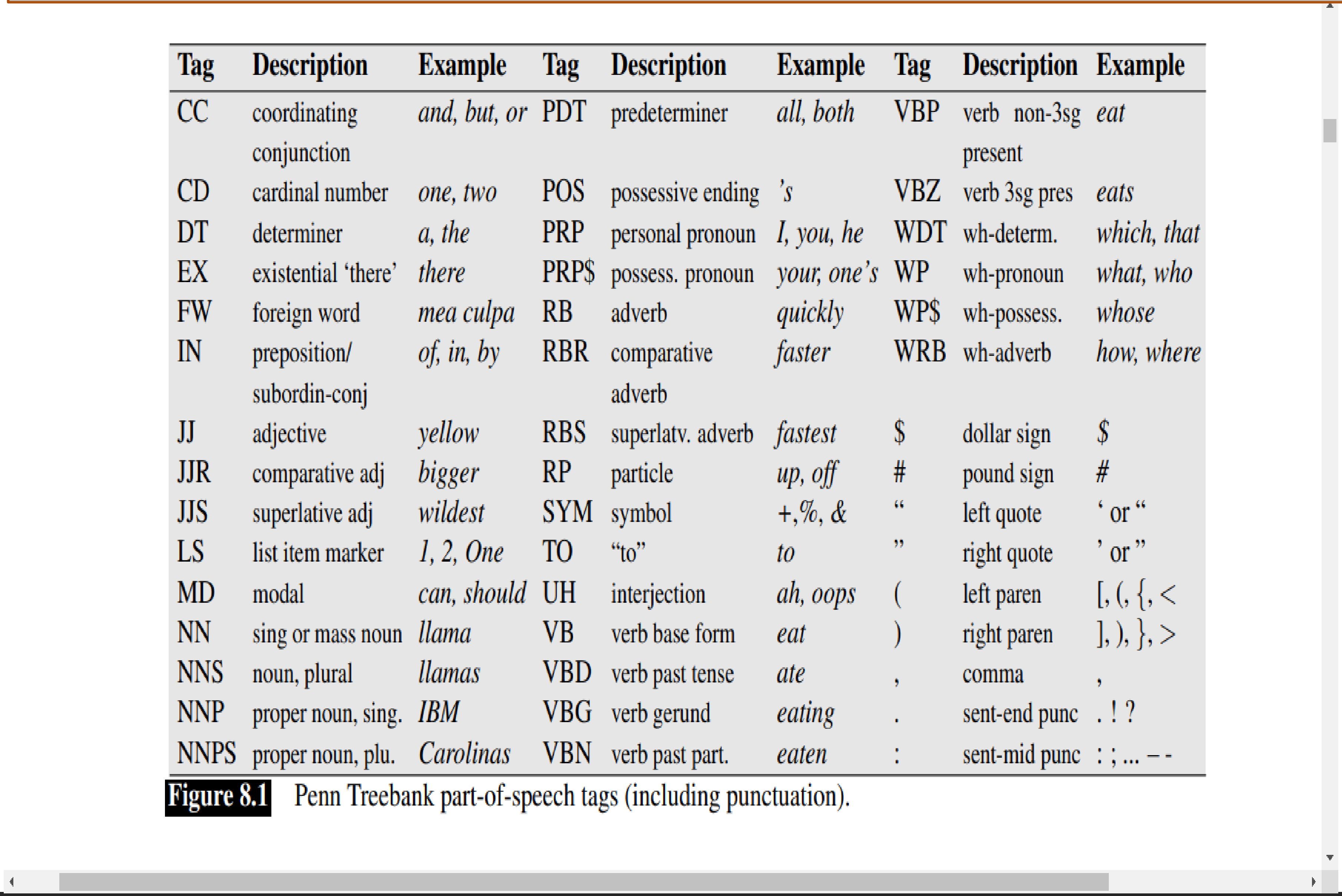
An important tagset for English is the 45-tag Penn Treebank tagset which has been used to label many corpora.

In such labelings, parts of speech are generally represented by placing the tag after each word, delimited by a slash (/). E.g.

- *The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.*

The e.g. shows the determiners *the* and *a*, the adjectives *grand* and *other*, the common nouns *jury, number*, and *topics*, and the past tense verb *commented*.

Penn Treebank POS Tags



Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

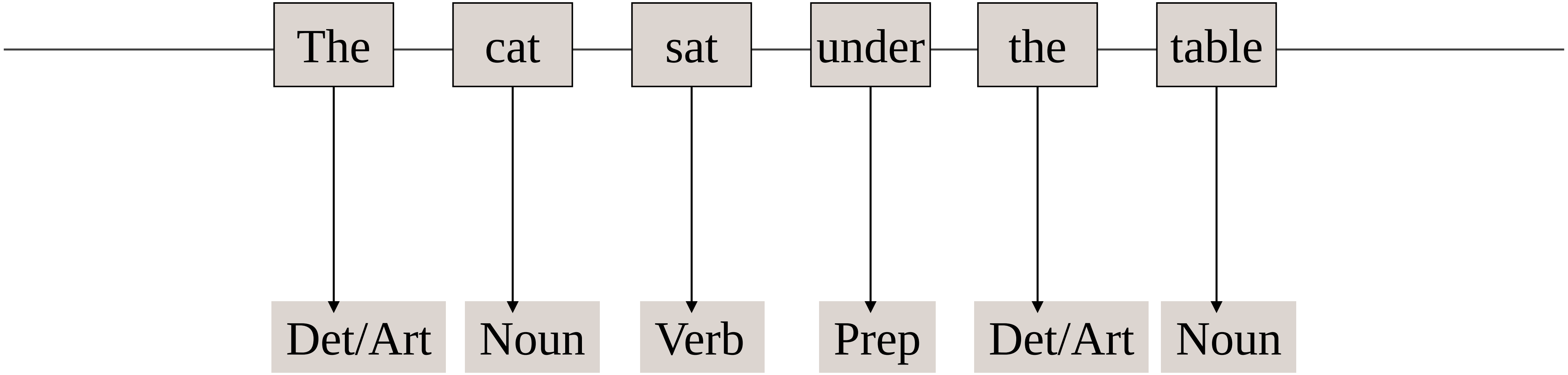
Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

Corpora labeled with parts of speech are crucial training (and testing) sets for statistical tagging algorithms.

Three main tagged corpora are consistently used for training and testing POS taggers for English:

- **Brown corpus** is a million words of samples from 500 written texts from different genres published in the US in 1961.
- **WSJ corpus** contains a million words published in the Wall Street Journal in 1989.
- **Switchboard corpus** consists of 2 million words of telephone conversations collected in 1990-1991.

POS TAGGING



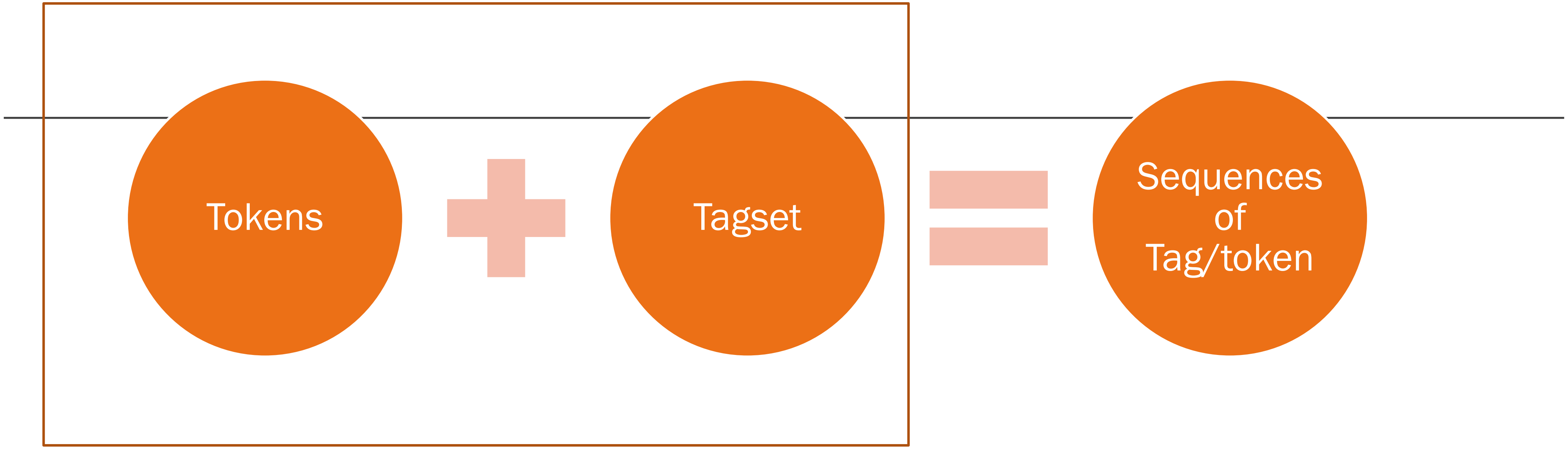
The cat sat under the table

Det/Art Noun Verb Prep Det/Art Noun

What is POS tagging?

- POS tagging is the process of assigning a part of speech or other syntactic class marker to each word in an input text.

Tagging Algorithm



Input:
Book that flight

Tagset:
Penn Treebank
(45 tagsets)

Output:
Book/VB *that*/DT *flight*/NN

Tagging is a disambiguation task

Words are ambiguous —have more than one possible part-of-speech

Issue of POS tagging: to determine the POS tag for a particular instance of a word.

E.g. *back:*

- *The back/JJ door*
- *On my back/NN*
- *Win her heart back/RB*

The goal of POS-tagging is to resolve these ambiguity, choosing the proper tag for the context.

Which of the following is not a possible interpretation of "*Fruit flies like a banana*"?

N = noun, V = verb, P = preposition, and DT = determiner.

a) *Fruit*/N *flies*/N *like*/V *a*/DT *banana*/N

b) *Fruit*/N *flies*/V *like*/IN *a*/DT *banana*/N

c) *Frui*t/N *flies*/N *like*/IN *a*/DT *banana*/N

d) None of the above (i.e., all the above are possible interpretations)

How common is tag ambiguity?

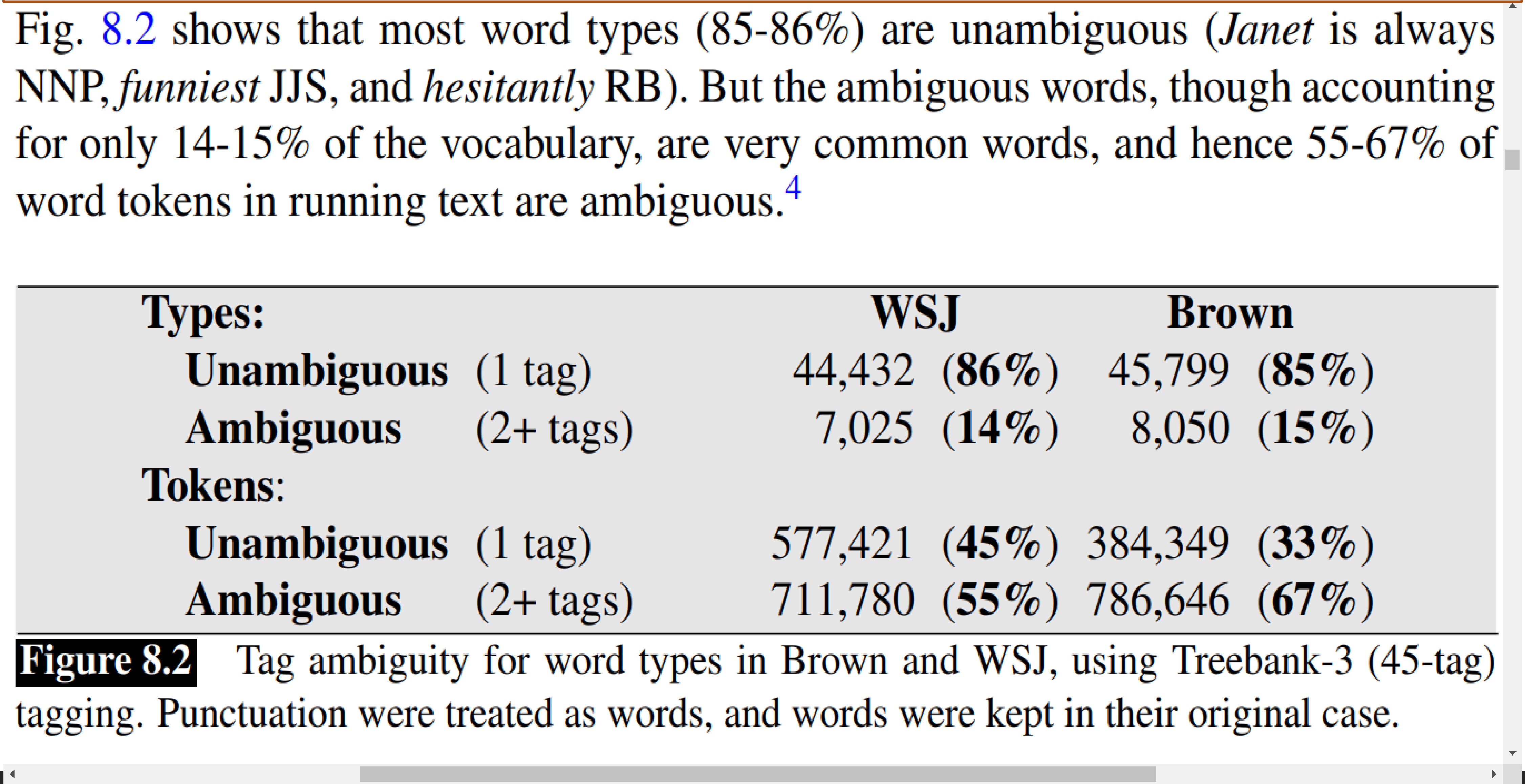


Fig. 8.2 shows that most word types (85-86%) are unambiguous (*Janet* is always NNP, *funniest* JJS, and *hesitantly* RB). But the ambiguous words, though accounting for only 14-15% of the vocabulary, are very common words, and hence 55-67% of word tokens in running text are ambiguous.[4]

Types:		**WSJ**	**Brown**
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Figure 8.2 Tag ambiguity for word types in Brown and WSJ, using Treebank-3 (45-tag) tagging. Punctuation were treated as words, and words were kept in their original case.

Some of the most ambiguous frequent words are:

- *that, back, down, put* and *set*;

To resolve ambiguous words, a simplistic baseline algorithm for POS tagging: *given an ambiguous word, choose the tag which is most frequent in the training corpus.*

A standard way to measure the performance of POS taggers is accuracy: the percentage of tags correctly labeled (matching human labels on a test set).

Tagging Algorithms

Two classes of tagging algorithms:

- Rule-based taggers
  - The earliest algorithms for automatically assigning POS were based on 2 stages:
    - Used dictionary to assign each word a list of potential POS.
    - Used large lists of hand-written disambiguation rules to winnow down this list to a single POS for each word.
- Probabilistic or stochastic taggers
  - HMM POS Tagging

HMM POS Tagging

Hidden Markov Model is a probabilistic sequence model:

- given a sequence of units, it computes a probability distribution over possible sequences of labels and chooses the best label sequence.

HMM is based on augmenting the Markov chain.

A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state. E.g. Weather prediction.

Consider a sequence of state variables $q_1, q_2, \ldots, q_i$. A Markov model embodies the Markov assumption on the probabilities of this sequence.

Markov Assumption:

$$P(q_i \mid q_1 \ldots q_{i-1}) = P(q_i \mid q_{i-1})$$

Components of HMM tagger:

- A probabilities: the probability of a tag occurring given the previous tag,

$$P(t_i \mid t_{i-1}) = C(t_{i-1}, t_i) \,/\, C(t_{i-1})$$

  - E.g. MD like *should* is very likely to be followed by a VB, like *go*, so we expect this probability to be high.

- B probabilities: the probability, given a tag, that it will be associated with a given word.

$$P(w_i \mid t_i) = C(t_i\,, w_i) \,/\, C(t_i)$$

  - E.g. A tagset MD is very likely to be associated with the word *go*.