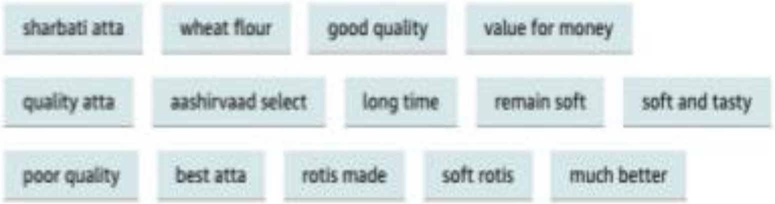
POS TAGGING

Advanced Text Pre-processing



sharbati atta | wheat flour | good quality | value for money

quality atta | aashirvaad select | long time | remain soft | soft and tasty

poor quality | best atta | rotis made | soft rotis | much better

- Extract nouns or adjectives or verbs or names of places or person names?
- Advanced Text Pre-processing NLP Tasks
  - POS Tagger – Parts of Speech Tagger
  - NER – Named Entity Recognizer
  - Parsers – Dependency and Constituency
  - Chunking
  - ..

AMRITA AHEAD Online

Why POS Tagger?

- Analyze Product Reviews:
  - Extract Product Descriptors:
    - Good, Nice, Bad, Useful
    - Grilled, Juicy, Spicy, Fresh
- These descriptors are adjectives you need to extract all adjectives.
- Solution: POS Tagger

AMRITA | AHEAD Online

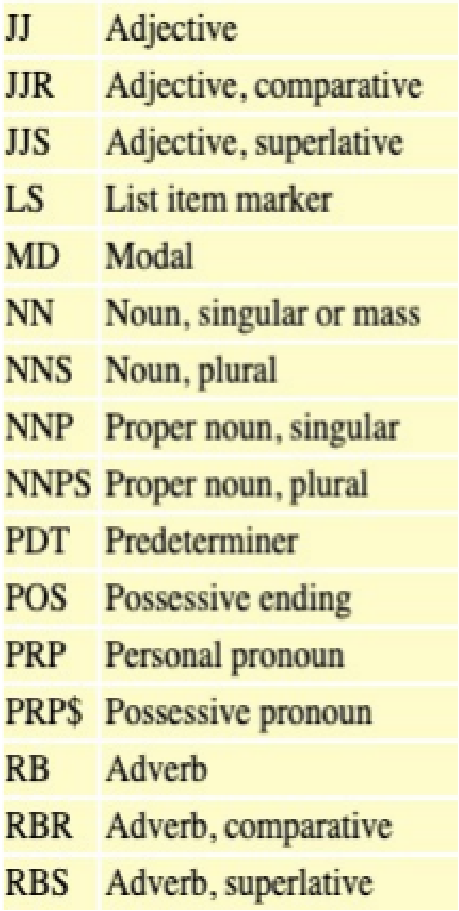
Parts-of-Speech (PoS) tagging

- PoS tagging is the process of tagging each word in sentences with their respective Parts of Speech – noun, verb.
- I am learning NLP.

I : Noun, am : Verb, learning: Verb, NLP: Noun

- Tagging is based on some Tag set : https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

AMRITA VISHWA VIDYAPEETHAM | AHEAD Online



JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative

[illegible]

Parts of Speech (POS)

- POS, word classes, or syntactic categories
- Why its important? → reveal a lot about a word and its neighbors.
- Parts of speech are useful features for
  - labeling named entities like people or organizations in information extraction
  - Generating syntactic structure of a given sentence used for parsing.
  - Machine Translation
  - Resolve ambiguity as words are ambiguous
    - Book me a flight.
    - Give me that book.

AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

Part-Of-Speech

- A POS - part of speech is a group of words that have common grammatical features.
  - **Noun** - The name of a person, place, thing, or idea
    - Book, pen, Amrita University, students
  - **Verb** - The action or being
    - Do, does, doing, read,
  - **Adjective** - This modifies or describes a noun or a pronoun
    - Beautiful, good, wonderful
  - **Adverb** - This modifies or describes a verb, adjective, or another adverb
    - Slowly, quickly, steadily

AMRITA VISHWA VIDYAPEETHAM AHEAD Online

Part-Of-Speech

- **Pronoun** - The word to be used in place of a noun
  - He, she, it
- **Preposition** - The word placed before a noun or pronoun to form a phrase modifying another word in the sentence
  - In, to, between
- **Conjunction** - This joins words, phrases, or clauses
  - And, but, so
- **Interjection** - A word used to express emotion
  - Oh!, Alas!

AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

Word classes

- Parts of speech can be divided into two broad .
- **Closed class**
  - Closed classes are those with relatively fixed membership, such as prepositions, pronouns (he, between, it)
  - Here word list is very short, occur frequently, and often have structuring uses in grammar.
- **Open class**
  - Open class POS tags can accept new words. (covidiot, google it, …)
  - Nouns, verbs, adjectives and adverbs are Open class

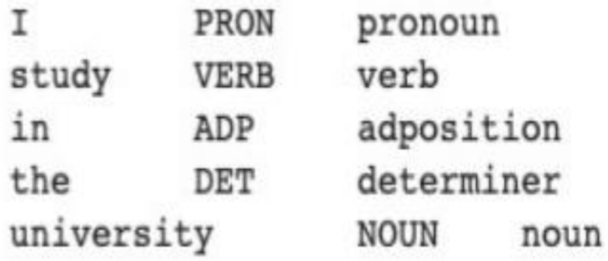
AMRITA VISHWA VIDYAPEETHAM AHEAD Online

POS TAGGER

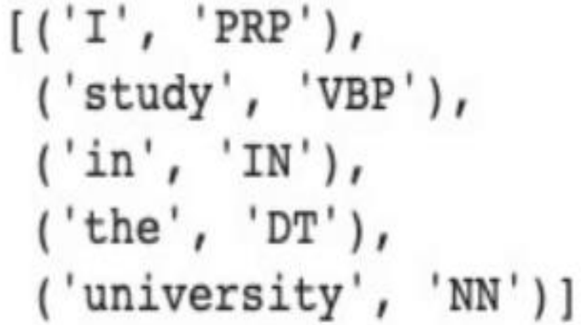
Parts-of-Speech (PoS) tagging

- PoS tagging is the process of assigning a part-of-speech tag / marker to each word in a given sentence.
- A POS tagging algorithm/model takes
  - is a sequence of (tokenized) words and a tagset, and
  - outputs is a sequence of tags one per token.
- Tag set : https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

AMRITA AHEAD Online



```
I           PRON     pronoun
study       VERB     verb
in          ADP      adposition
the         DET      determiner
university           NOUN     noun
```

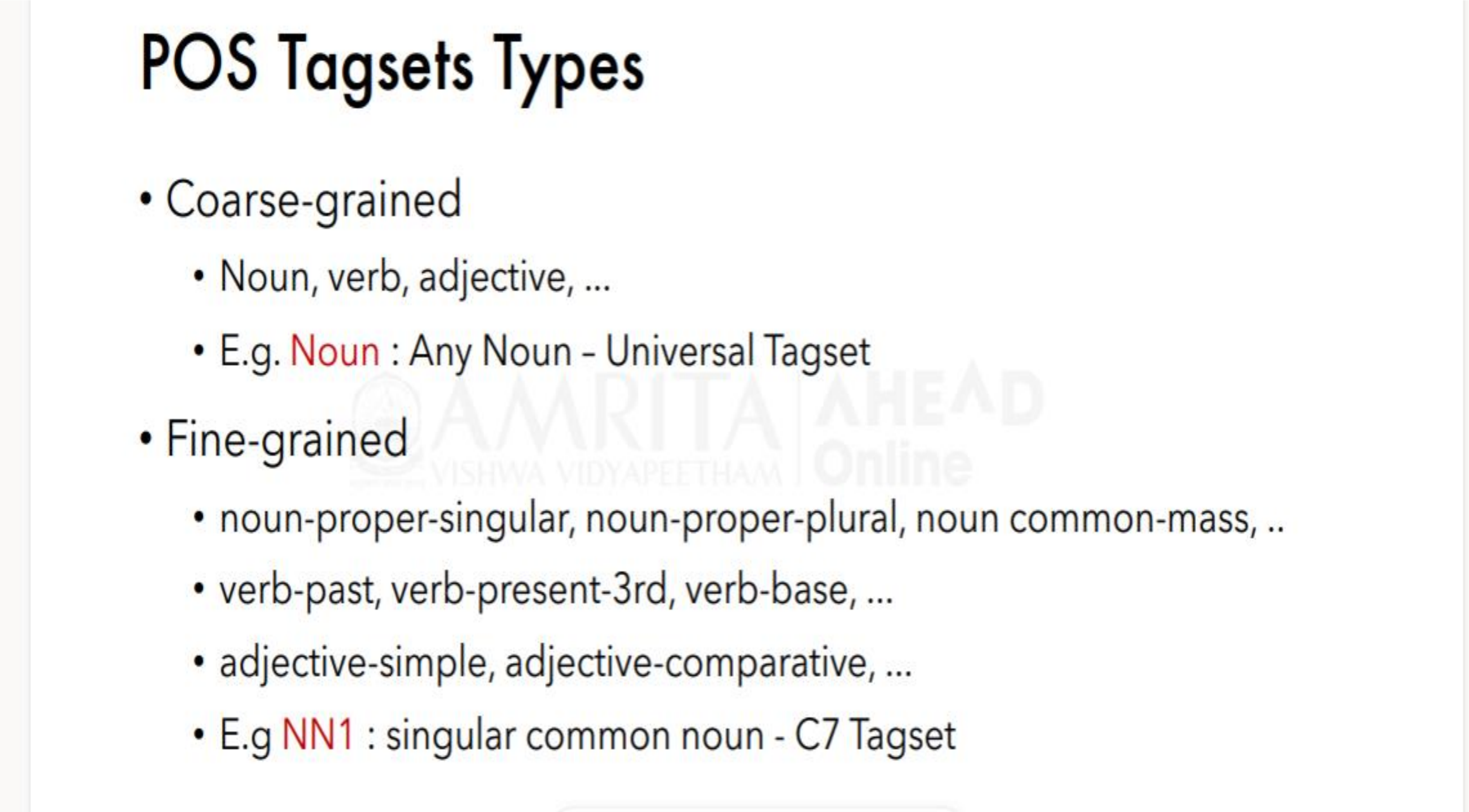


```
[('I', 'PRP'),
 ('study', 'VBP'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('university', 'NN')]
```

POS Tagsets Types

- POS tags are used to annotate/tag words and mark their POS.
- E.g : Book/Noun, Saw/Verb, Good/ADJ,
- There are many types of parts of speech tagsets
  - Coarse-grained – general marking is done
    - Noun, NN,
  - Fine-grained – specific and detailed tagging is done
    - verb-present-3rd, common-noun-plural

AMRITA AHEAD Online



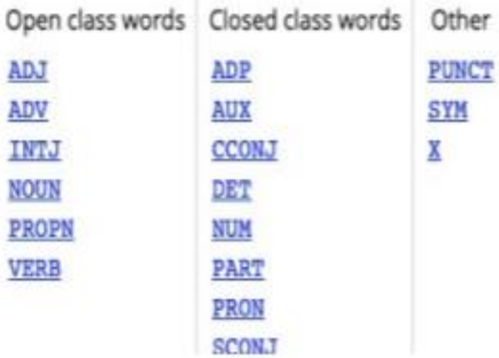
POS Tagsets Types

- Coarse-grained
  - Noun, verb, adjective, ...
  - E.g. Noun : Any Noun - Universal Tagset
- Fine-grained
  - noun-proper-singular, noun-proper-plural, noun common-mass, ..
  - verb-past, verb-present-3rd, verb-base, ...
  - adjective-simple, adjective-comparative, ...
  - E.g NN1 : singular common noun - C7 Tagset

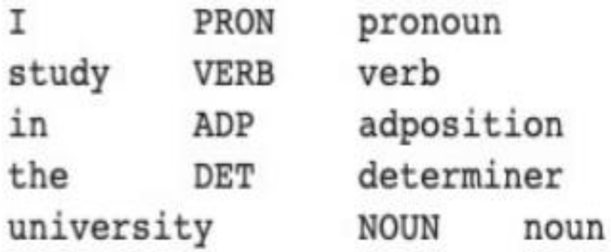
AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

POS Tagsets

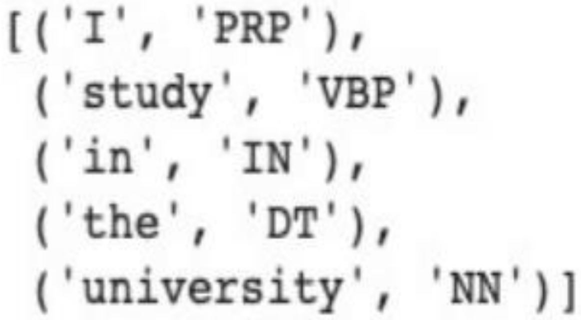
- Brown tagset (87 tags) - Brown corpus
- C5 tagset (61 tags)
- C7 tagset (146 tags!)
- Penn TreeBank (45 tags) - most used
- Universal Tag set (15 tags)
  - Coarse-grained



Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	



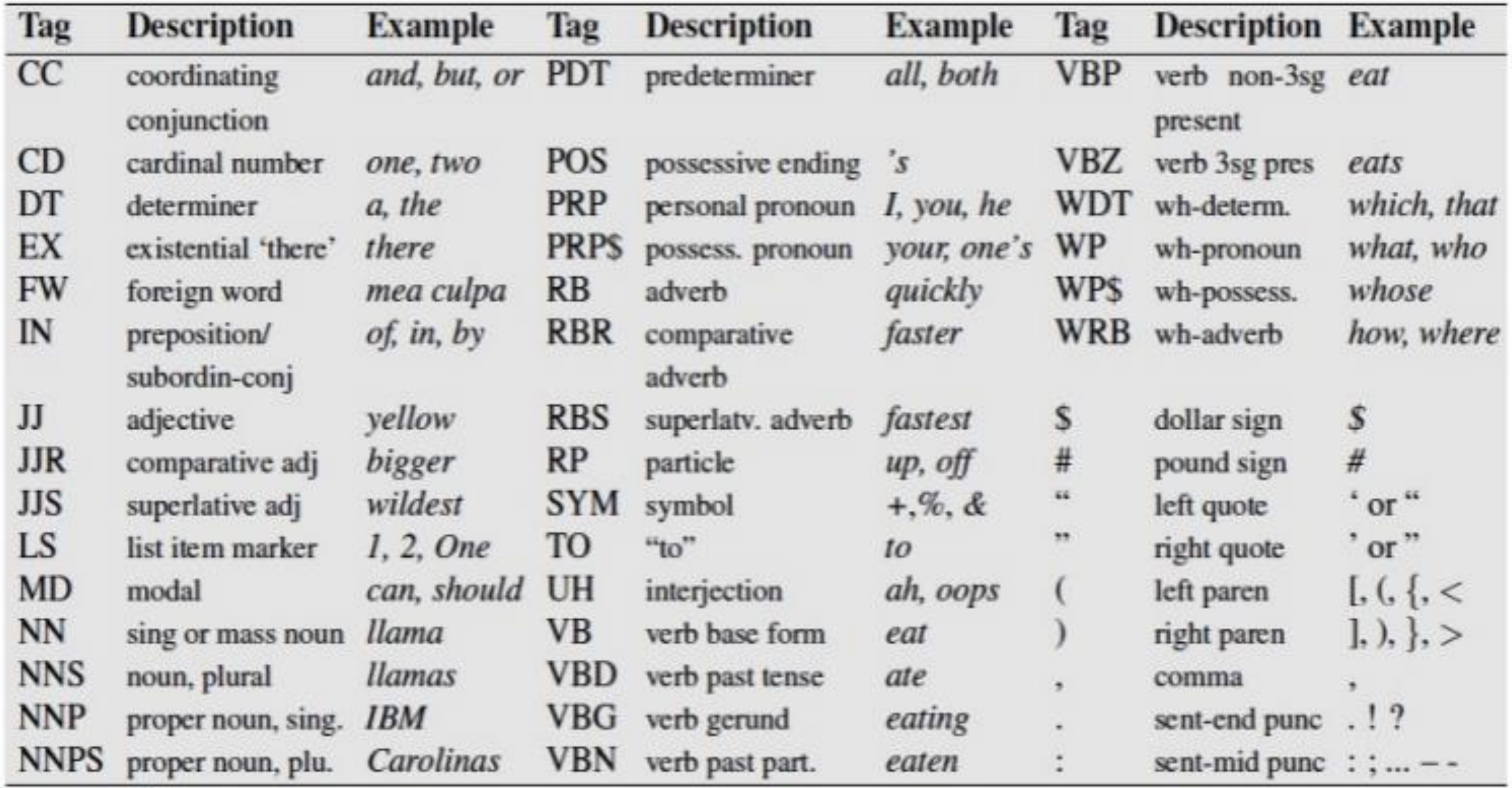
```
I           PRON      pronoun
study       VERB      verb
in          ADP       adposition
the         DET       determiner
university            NOUN      noun
```



```
[('I', 'PRP'),
 ('study', 'VBP'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('university', 'NN')]
```

AMRITA | AHEAD VISHWA VIDYAPEETHAM | Online

Tagging words – Penn TreeBank Tagset



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).

Tagging Algorithms

- HMM : Hidden Markov Model
- MEMM : Maximum Entropy Markov Models
- Machine Learning
- Neural networks
- Rule-based algorithms

AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

nltk.tag Module

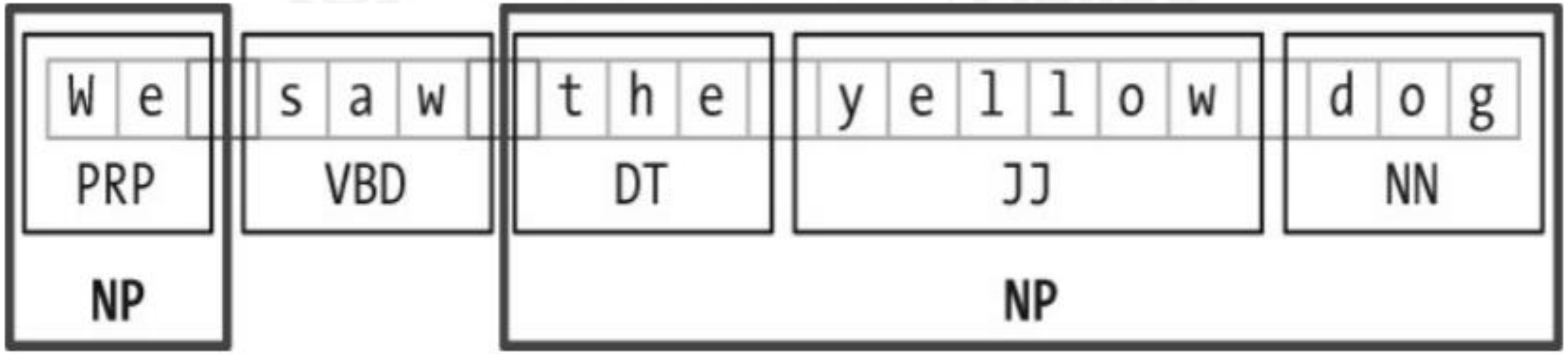
- PerceptronTagger
- StanfordPOSTagger
- HMMTagger
- HunposTagger

AMRITA VISHWA VIDYAPEETHAM | AHEAD Online

Chunking and Chinking

- Chunking : **Groups Segments** and labels multitoken sequences.
- **Removing specific chunks** after chunking → chinking.

AMRITA AHEAD Online



Named Entity Recognition

- Task finding proper names or named named entity entities in a text
- **extracting important entities**, such as person names, place names, and organization names, from some given text.

Amrita Vishwam **studied in** Amrita Vishwa Vidyapeetham **in the year** 2006**. She stayed in** Kollam**.**

- Person Date Organization Location

AMRITA VISHWA VIDYAPEETHAM AHEAD Online