POS Tagging with NLTK

Eng. Fatma

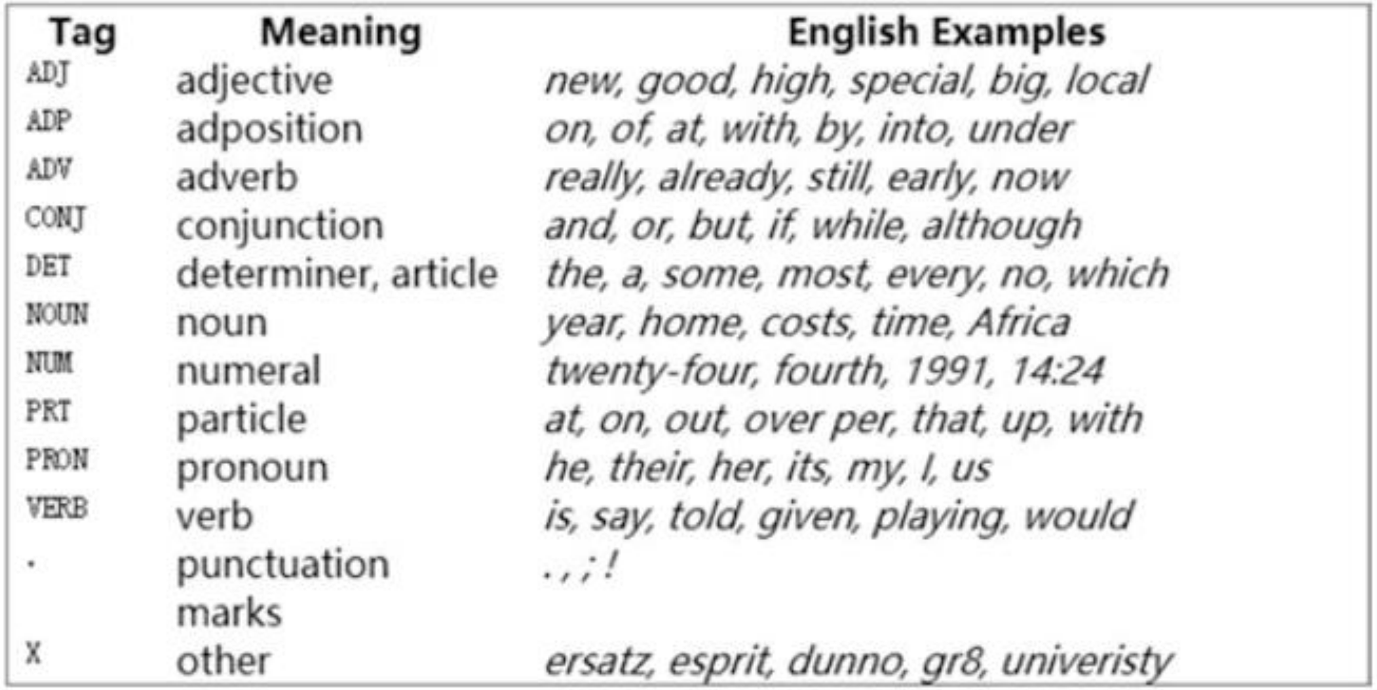
What Is POS Tagging?

- Part-of-Speech (POS) refers to words categorization process in a sentence/utterance into specific syntactic or grammatical functions.
- There are 9 major POS in English: Nouns, Pronouns, Adjectives, Verbs, Prepositions, Adverbs, Determiners, interjection, and Conjunctions.
- POS tagging is to assign POS tags into each word token in the sentence/utterance.

Universal POS Tagset

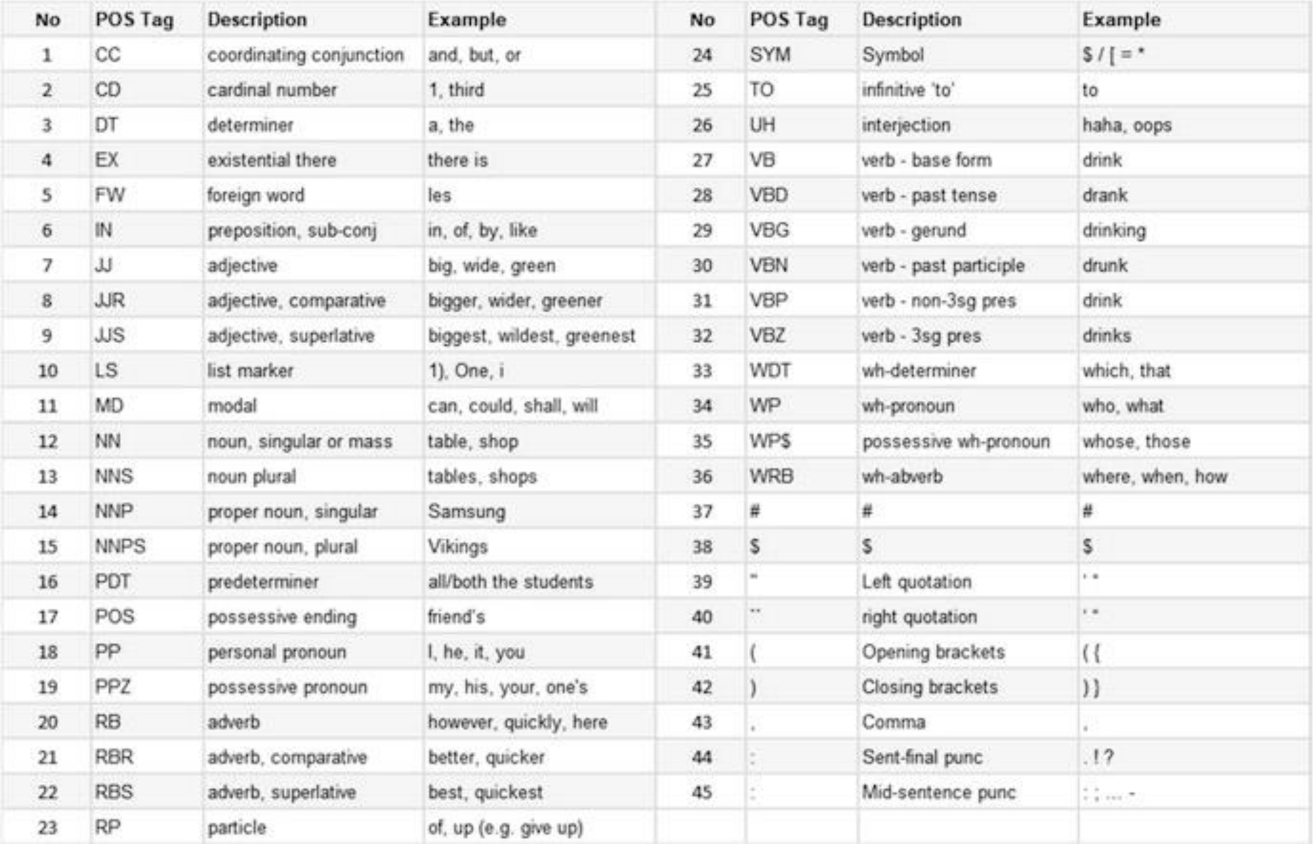
- A tagset consists of 12 universal POS categories which is constructed to facilitate future requirements for unsupervised induction of syntactic structure.
- When is combined with original treebank data, this universal tagset and mapping produce a dataset consisting of common POS in 22 languages.

Universal POS Tagset



Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

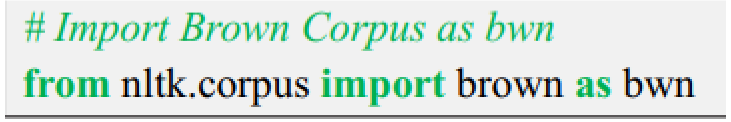
PENN Treebank Tagset (English and Chines)



No	POS Tag	Description	Example	No	POS Tag	Description	Example
1	CC	coordinating conjunction	and, but, or	24	SYM	Symbol	$ / [= *
2	CD	cardinal number	1, third	25	TO	infinitive 'to'	to
3	DT	determiner	a, the	26	UH	interjection	haha, oops
4	EX	existential there	there is	27	VB	verb - base form	drink
5	FW	foreign word	les	28	VBD	verb - past tense	drank
6	IN	preposition, sub-conj	in, of, by, like	29	VBG	verb - gerund	drinking
7	JJ	adjective	big, wide, green	30	VBN	verb - past participle	drunk
8	JJR	adjective, comparative	bigger, wider, greener	31	VBP	verb - non-3sg pres	drink
9	JJS	adjective, superlative	biggest, wildest, greenest	32	VBZ	verb - 3sg pres	drinks
10	LS	list marker	1), One, i	33	WDT	wh-determiner	which, that
11	MD	modal	can, could, shall, will	34	WP	wh-pronoun	who, what
12	NN	noun, singular or mass	table, shop	35	WP$	possessive wh-pronoun	whose, those
13	NNS	noun plural	tables, shops	36	WRB	wh-abverb	where, when, how
14	NNP	proper noun, singular	Samsung	37	#	#	#
15	NNPS	proper noun, plural	Vikings	38	$	$	$
16	PDT	predeterminer	all/both the students	39	"	Left quotation	' "
17	POS	possessive ending	friend's	40	''	right quotation	' "
18	PP	personal pronoun	I, he, it, you	41	(	Opening brackets	({
19	PPZ	possessive pronoun	my, his, your, one's	42	)	Closing brackets	) }
20	RB	adverb	however, quickly, here	43	,	Comma	,
21	RBR	adverb, comparative	better, quicker	44	:	Sent-final punc	. ! ?
22	RBS	adverb, superlative	best, quickest	45	:	Mid-sentence punc	: ; ... -
23	RP	particle	of, up (e.g. give up)				

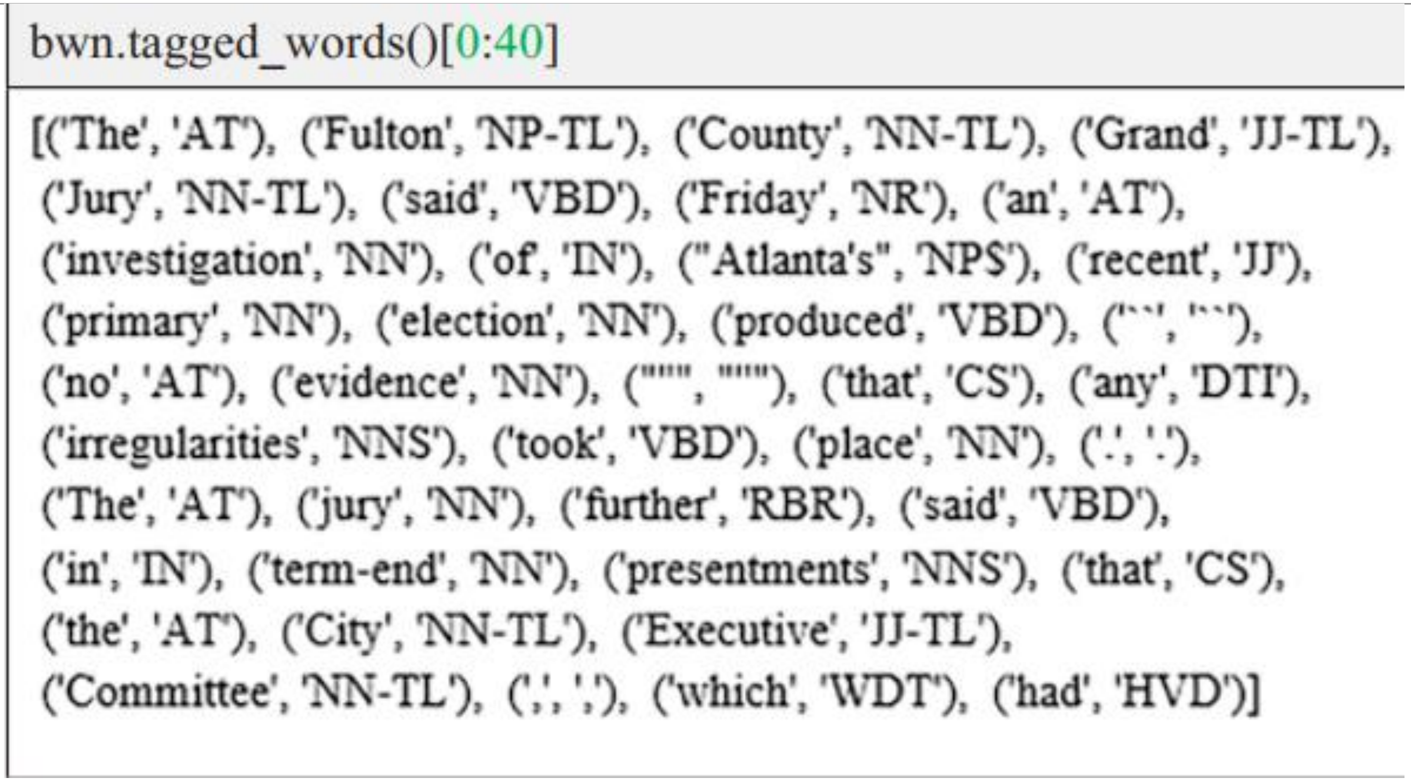
PENN Treebank Tagset (English and Chines)

- NLTK provides direct mapping from tagged corpus such as Brown Corpus (NLTK 2022) to universal tags for implementation, e.g. tags VBD (for past tense verb) and VB (for base form verb) map to VERB only in universal tagset.



```
# Import Brown Corpus as bwn
from nltk.corpus import brown as bwn
```

PENN Treebank Tagset (English and Chines)



```
bwn.tagged_words()[0:40]
```

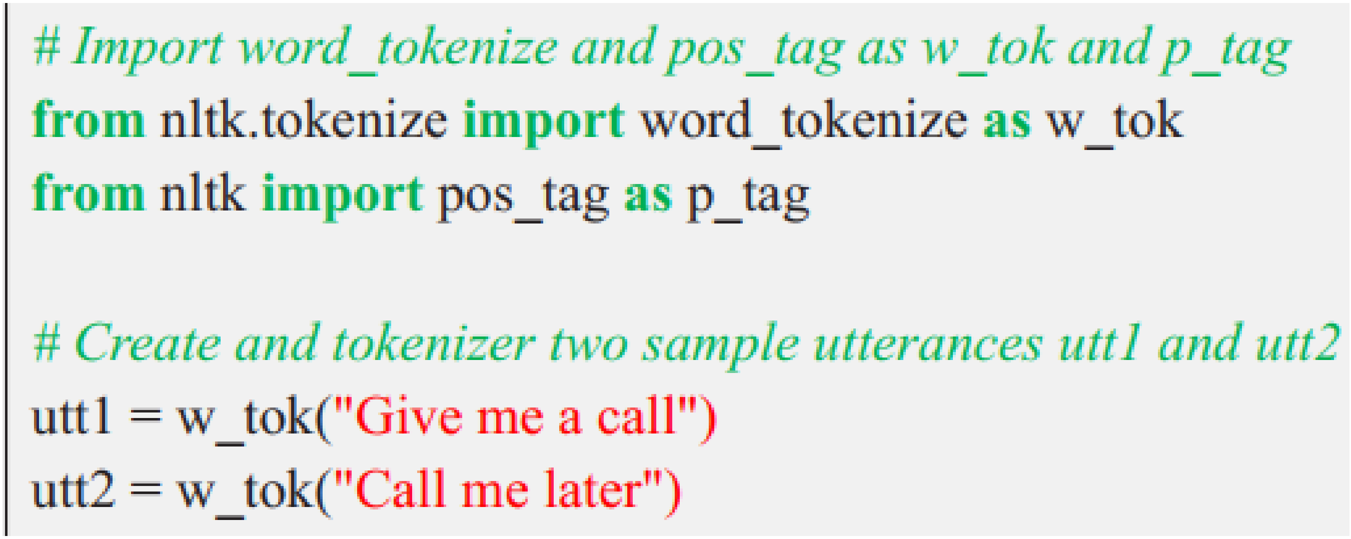
```
[('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'),
 ('Jury', 'NN-TL'), ('said', 'VBD'), ('Friday', 'NR'), ('an', 'AT'),
 ('investigation', 'NN'), ('of', 'IN'), ("Atlanta's", 'NP$'), ('recent', 'JJ'),
 ('primary', 'NN'), ('election', 'NN'), ('produced', 'VBD'), ('``', '``'),
 ('no', 'AT'), ('evidence', 'NN'), ("''", "''"), ('that', 'CS'), ('any', 'DTI'),
 ('irregularities', 'NNS'), ('took', 'VBD'), ('place', 'NN'), ('.', '.'),
 ('The', 'AT'), ('jury', 'NN'), ('further', 'RBR'), ('said', 'VBD'),
 ('in', 'IN'), ('term-end', 'NN'), ('presentments', 'NNS'), ('that', 'CS'),
 ('the', 'AT'), ('City', 'NN-TL'), ('Executive', 'JJ-TL'),
 ('Committee', 'NN-TL'), (',', ','), ('which', 'WDT'), ('had', 'HVD')]
```

Applications of POS Tagging

- POS tagging is commonly used in many NLP applications ranging from Information Extraction (IE), Named Entity Recognition (NER) to Sentiment Analysis and Question-&-Answering systems.

POS Tagging with nltk: Example

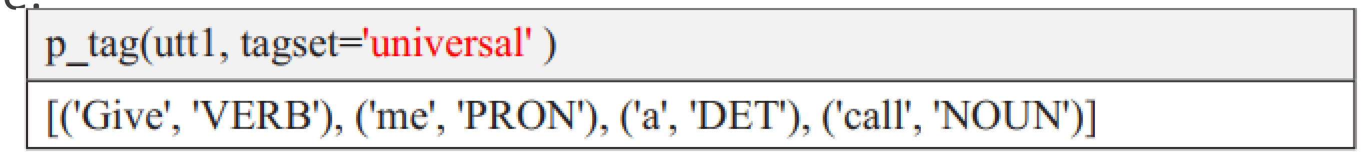
- Example:



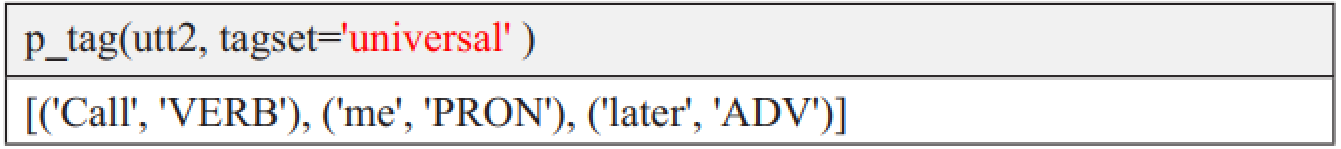
```
# Import word_tokenize and pos_tag as w_tok and p_tag
from nltk.tokenize import word_tokenize as w_tok
from nltk import pos_tag as p_tag

# Create and tokenizer two sample utterances utt1 and utt2
utt1 = w_tok("Give me a call")
utt2 = w_tok("Call me later")
```

POS Tagging with nltk: Example

- Example:



```
p_tag(utt1, tagset='universal' )
[('Give', 'VERB'), ('me', 'PRON'), ('a', 'DET'), ('call', 'NOUN')]
```



```
p_tag(utt2, tagset='universal' )
[('Call', 'VERB'), ('me', 'PRON'), ('later', 'ADV')]
```

POS Tagging with nltk: Example

- Notes:

1. The word call is a noun in text 1 and a verb in text 2.

2. POS tagging is used to identify a person, a place, or a location, based on the

Tags.

POS Tagging with nltk: Exercise

- NLTK also provides a classifier to identify such entities in text as shown in the following code.

- Try to use nltk classifier on the previous example.