Natural Language Processing

Dr. Karen Mazidi

Part One: Foundations

Topics

- Linguistics

Quizzes

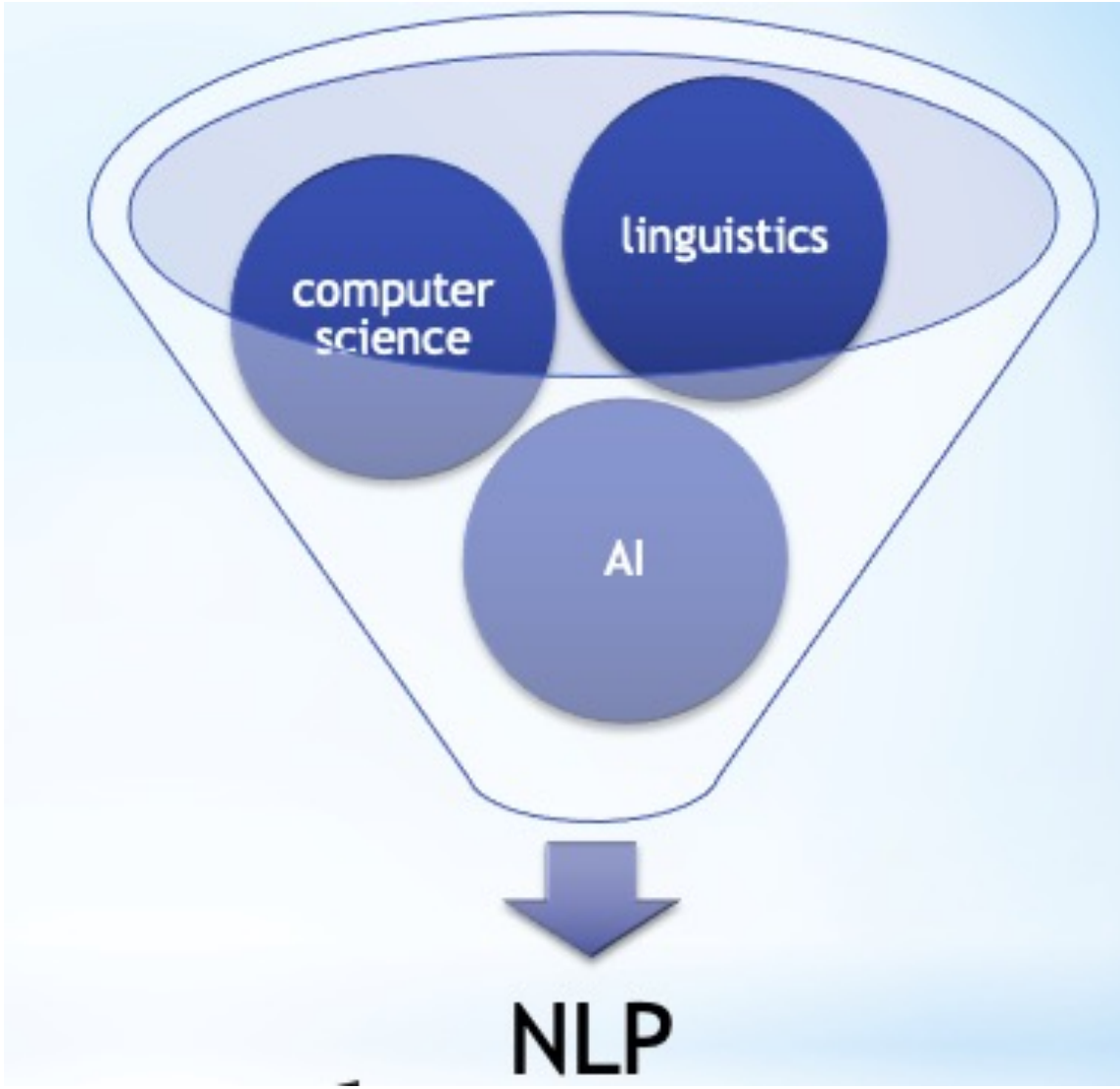
- Quiz: Linguistics Terminology

Homework

- not on this material

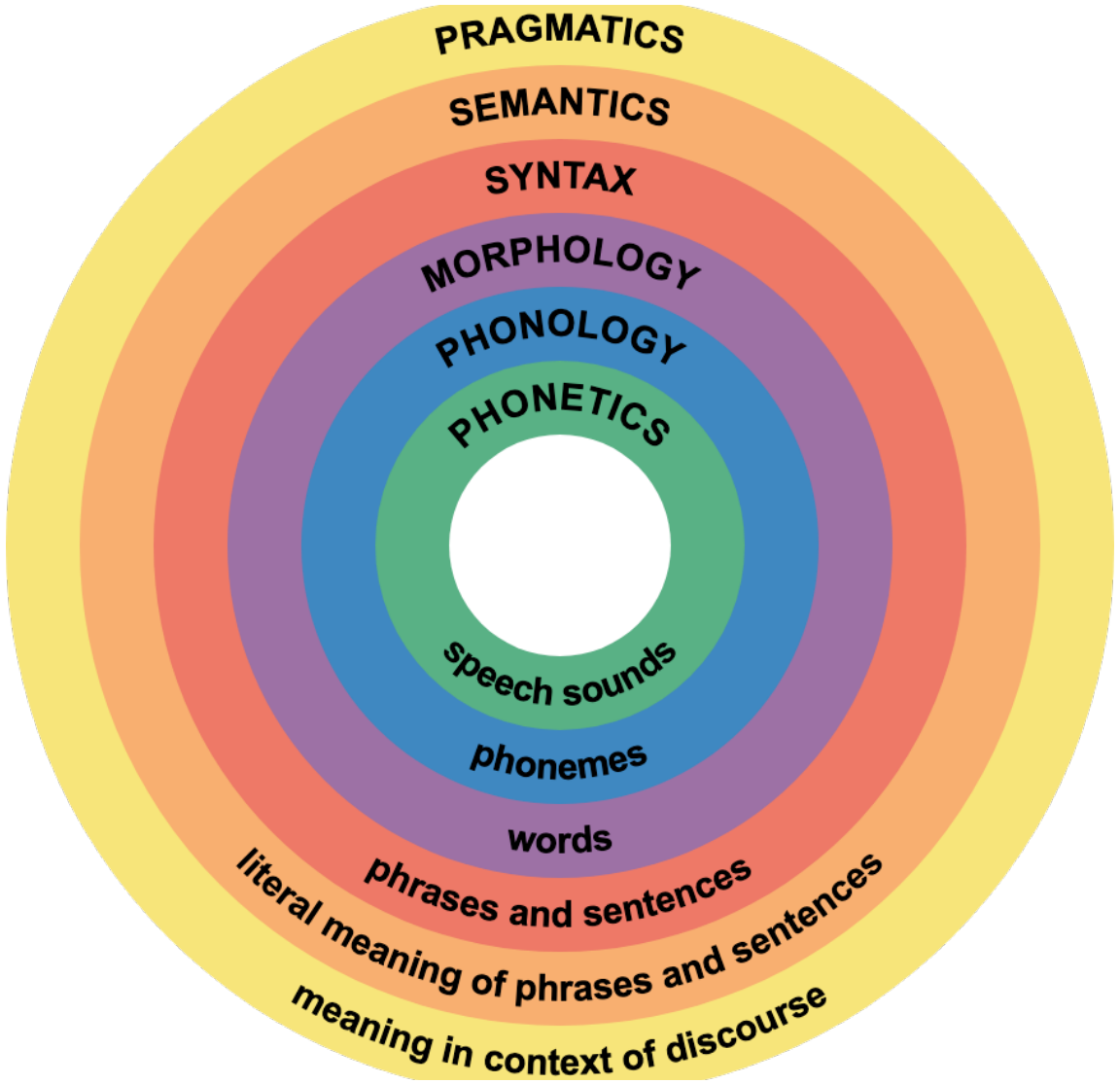
Agenda

- Linguistic terms
- Linguistic concepts
- What linguists do



Linguistics

- The study of human languages
  - How produced
  - How language evolves
  - How children learn language
  - Relationship between language and thought



Language levels

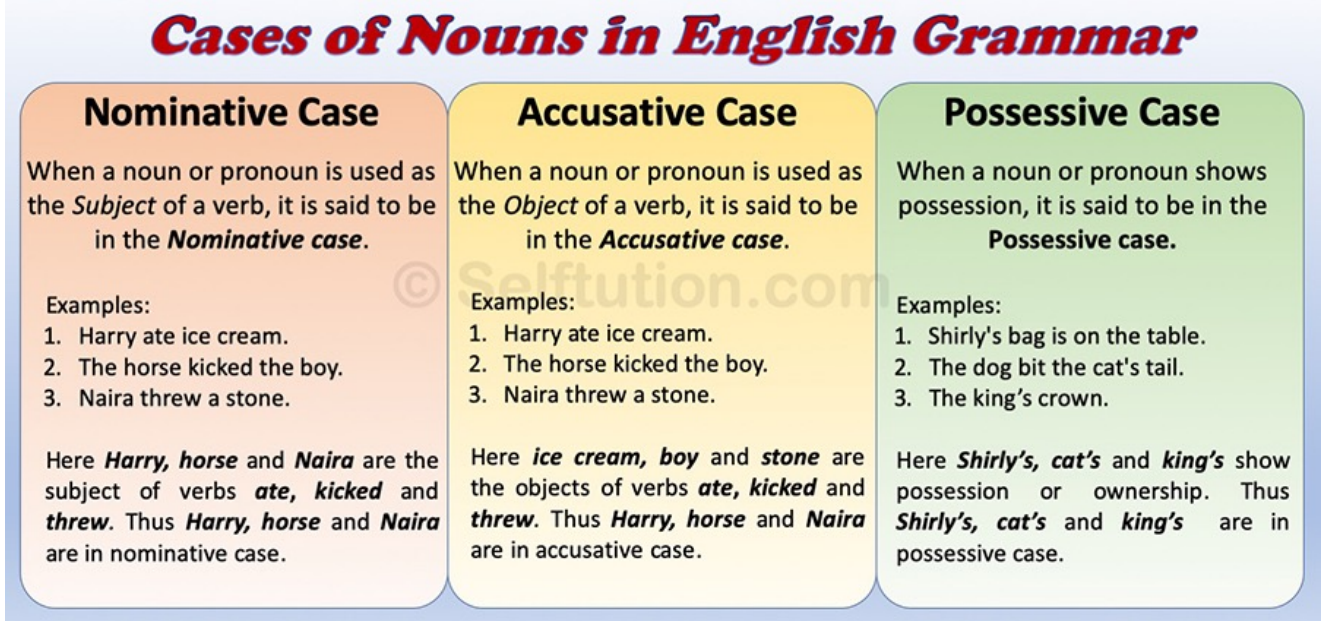
- phonology – studies the sounds of a language
  - ex: English has about 45 phonemes that are combined into syllables to form words
- morphology – studies the way words change shape
  - ex: run, running, ran
  - affixes: pre- and post- change word meaning
  - morphemes – components of words that carry meaning, like un-chang-ing-ly
- orthography – written equivalent of morphemes
- lexicon – set of lemmas (base form of words) of a language
- syntax – how words are grammatically combined into sentences
- semantics – the meaning of sentences
- pragmatics – rules of discourse, context, sarcasm, etc.

Word categories

- POS parts of speech
  - Relatively stable across languages
  - Noun, pronoun, verb, adverb, adjective, etc.
- Linguists divide words into two categories:
  - Open – open to addition
    - Nouns, verbs, adjectives, adverbs
    - Content words
  - Closed – only change over centuries
    - Preposition, conjunction, articles, etc.
    - Function words aka glue words

Nouns

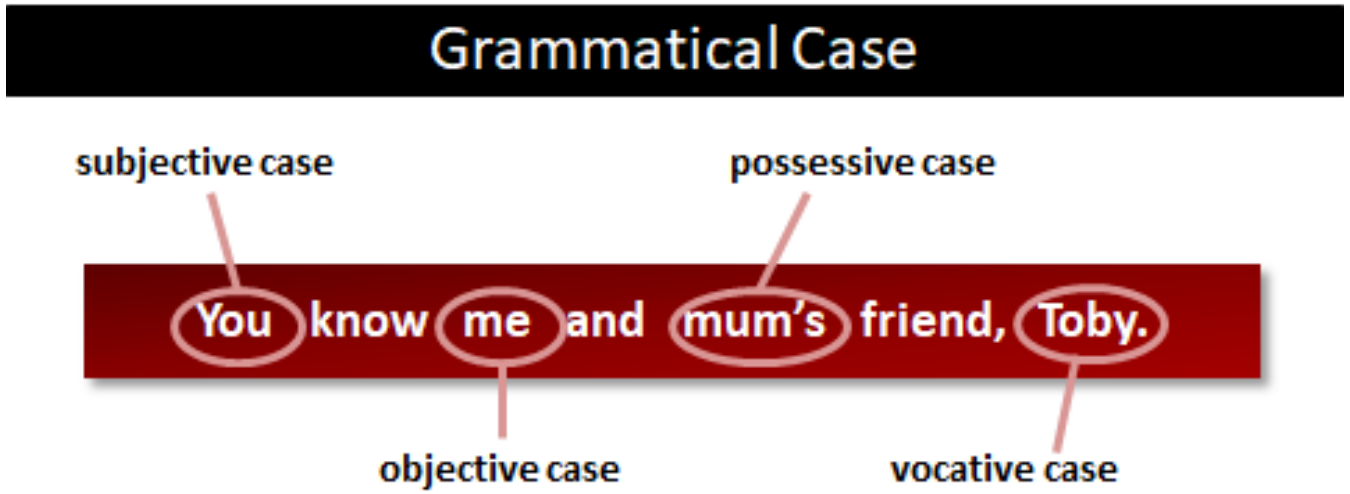
- People, places, things, abstract ideas, all called entities in NLP
- Nouns inflect, but vary by language:
  - Number: singular, plural
  - Gender: m, f, neutral
  - Common, proper: girl, Julia
  - Case: vestiges of case in English



Pronouns

- categories
- inflect by case

- personal – vary by number and gender
  - he, him, they, you
- possessive – my, mine, their
- reflexive – myself, herself
- demonstrative – this, that, those, these
- indefinite – one, someone, few
- some sources have more categories



Verbs

- describe action or state
- inflection varies by language
- conjugate by 1st, 2nd, 3rd person singular, plural
- tense
- aspect, often temporal: progressive, perfect, …
- modality: modal verbs can, should, should,…
- voice: active, passive
- negation: English not; other languages inflectional

English tenses

		Tenses			
		Morphological		With auxiliaries	
		Present	Past	Future	Future-in-the-past
Aspects	Simple	*go(es)*	*went*	*will go*	*would go*
	Continuous	*am/is/are going*	*was/were going*	*will be going*	*would be going*
	Perfect	*have/has gone*	*had gone*	*will have gone*	*would have gone*
	Perfect continuous	*have/has been going*	*had been going*	*will have been going*	*would have been going*

Special Verbs

- copula verb – be: He is handsome
- auxiliary verbs – do, be, have
- light verbs: take, make, etc.
  - take a nap, make up with a friend

Adjectives

Modify nouns

- describe number
- inflectional: number, gender, etc
- often derived from nouns
  - amateur -> amateurish
  - trend -> trendy
  - hate -> hateful

Adverbs

Modify verbs

- often derived from adjectives:
  - swift -> swiftly
  - odd -> oddly

Other POS

- determiners: a, the, …
- prepositions: on, in, under, …
- particles: take up a hobby
- conjunction: and, but, or, although,
- coordinating: peas and carrots
- subordinating: He hates peas because they are green.
- existential ‘there’: There are two issues.
- negation: no, not, never

wh-words

aka interrogative words: who, what, where …

- can act as adverbs:
  - When/how/where did it happen?
- can act as pronouns:
  - Who ate the cookie? What sound was that?
- can act as determiners:
  - Which dog bit you?

Classifying languages

- no universal agreement
- commonly mentioned classifications:
- Indo-European, Romance, Semitic
- linguists tend to classify languages by morphological features:
- analytical languages like English, French, Japanese, use a lot of function words to convey meaning: The book is on that table.
- inflective languages like many Slavic languages and Arabic use affixes to convey meaning.
- agglutinative languages like Turkish and Hungarian combine morphemes

Agglutinative languages

- The Black Book by Orhan Pamuk

- Translator's notes:

  "Apparently they were inside their houses"

  - A single word in Turkish

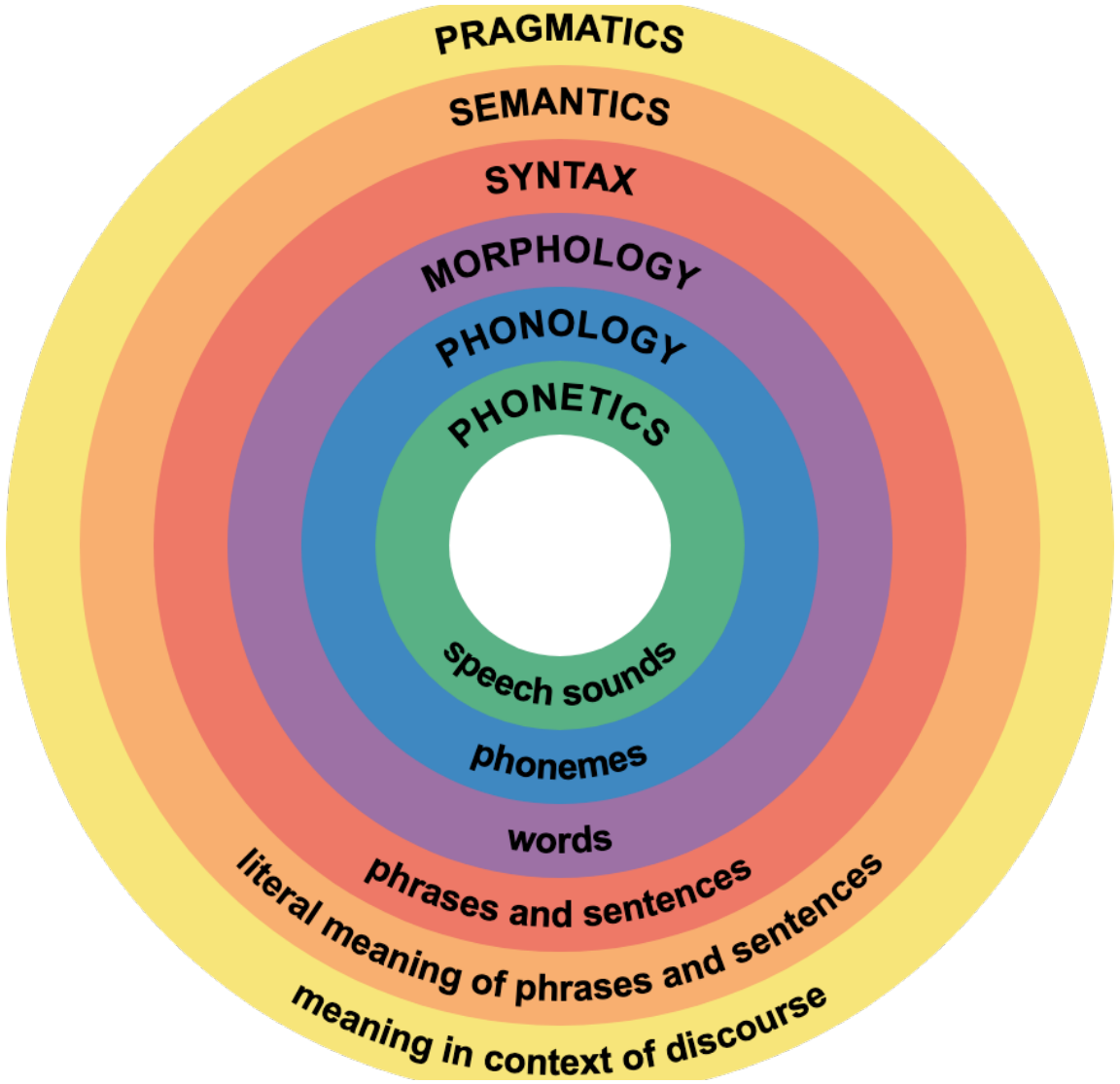
- Great history of Hungarian language:
- https://www.youtube.com/watch?v=ikODMvw76j4

Language Origins

- linguist Noam Chomsky and others argue that language is innate like learning to walk
  - good read: Steven Pinker, The Language Instinct
- Daniel Everett, field linguist, argues that communication and language are culturally transmitted, and became incrementally more complex through evolution, starting as far back as Homo Erectus
  - good read: How Language Began

Levels of language

- Surface: tokens, morphology
  - John ate (eat:past) a dog
- A little deeper: syntax
  - Subject verb determiner direct-object
- Meaning: semantics
  - Hot dog? 4-legged?
- Pragmatics
  - Ew!!!

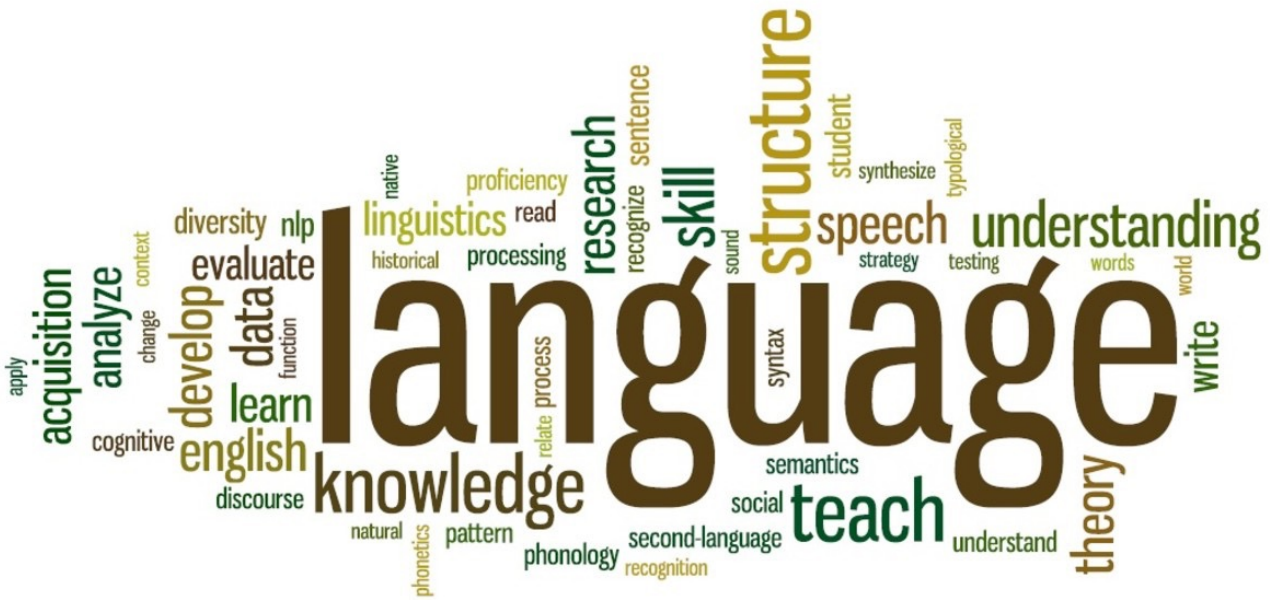


Natural Language Processing

What we look at . . .

What can computers ‘understand’

- We have NLP tools and techniques to explore:
  - words
  - parts of words
  - sentences
  - documents



Words

- tokens – meaning units (words, punctuation …)
- lemma - lexical unit
  - typically the base form of a word
  - the lemma of running is run
  - think of it as the form of the word you would look up in a dictionary
- stem – a stemmed word is a word cut down to a base form (via rules)
  - beauti – stem of beauty, beautiful, …
- word sense – the particular meaning a word has, gleaned from word meaning and context
  - bank of the river
  - get cash from the bank

POS (part of speech)

- we often need to know a word's part of speech
- challenging because the same word can have a different POS depending upon role in sentence:
  - I sense danger.
  - My spidey sense is tingling.

POS from NLTK

```
import nltk
from nltk import word_tokenize

sents = ['I sense danger.', 'My spidey sense is tingling.']
for s in sents:
    print('\n', s)
    tokens = word_tokenize(s)
    print('tokens: ', tokens)
    pos_tags = nltk.pos_tag(tokens)
    print('pos: ', pos_tags)
```

```
 I sense danger.
tokens:  ['I', 'sense', 'danger', '.']
pos:  [('I', 'PRP'), ('sense', 'VBP'), ('danger', 'NN'), ('.', '.')]

 My spidey sense is tingling.
tokens:  ['My', 'spidey', 'sense', 'is', 'tingling', '.']
pos:  [('My', 'PRP$'), ('spidey', 'NN'), ('sense', 'NN'), ('is', 'VBZ'), ('tingling', 'VBG'), ('.', '.')]
```

POS from Penn Treebank

- John broke the window.
  - John/NNP
  - broke/VBD
  - the/DT
  - window/NN
  - ./.
- The Penn Treebank is a human-annotated corpus of text, including the Brown corpus, the Wall Street Journal corpus, etc.
- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRPS	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Stop words

- function words like ‘the’, ‘is’, ‘at’ occur so often that we sometimes want to eliminate these stop words from our text

```
from nltk import word_tokenize
from nltk.corpus import stopwords

tokens = word_tokenize("All his exes live in Texas and that is why"
                       " he lives in Tennessee")
print('original: ', tokens)
stop_words = set(stopwords.words('english'))
words_filtered = [x for x in tokens if x not in stop_words]
print('filtered: ', words_filtered)
```

```
original:  ['All', 'his', 'exes', 'live', 'in', 'Texas', 'and', 'that', 'is', 'why', 'he', 'lives', 'in', 'Tennessee']
filtered:  ['All', 'exes', 'live', 'Texas', 'lives', 'Tennessee']
```

Morpheme

- a morpheme is a minimal meaning-bearing unit in a language
  - ex: base form (stem) believe
  - affixes (suffix, prefix, infix) un-, -able, -ly
- morphology is the study of the structure and formation of words
- important for many tasks such as machine translation, information retrieval, POS tagging, etc.

morphology

- morphemes combine to make words:

- inflection: clean -> cleaning (verb)
- derivation: clean -> cleaning (noun)
- compounding: firetruck
- cliticization: I’ve

Figurative speech

- Love is not a bed of roses.
- The well is dry as a bone.
- He kicked the bucket.

zeugma

- A literary or rhetorical device in which a word, usually a verb, extends to several phrases, example from Francis Bacon:

  Histories make men wise; poets, witty; the mathematics, subtle; natural philosophy, deep, moral, grave; logic and rhetoric, able to contend.

- Another example. From Star Trek:

  You are free to execute your laws, and your citizens, as you see fit.

Summary

- Some linguistic terminology that NLP practitioners should know, but we are not expected to know as much as linguists

To Do

- Quiz on Linguistic terms
- Homework questions?
- Next class: Part 2 Words