UNIVERSITÀ DEGLI STUDI DI NAPOLI
PARTHENOPE

Natural Language Processing

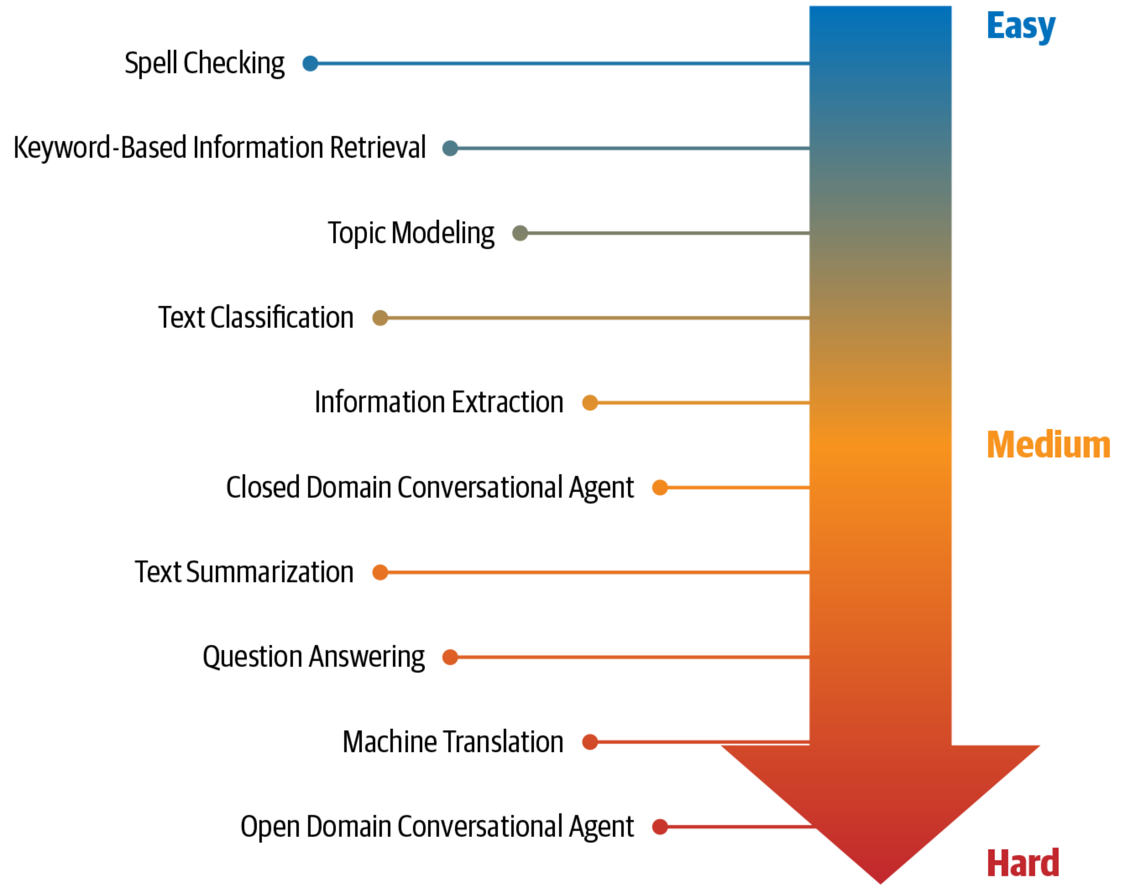
Elements of Linguistics

LESSON 2

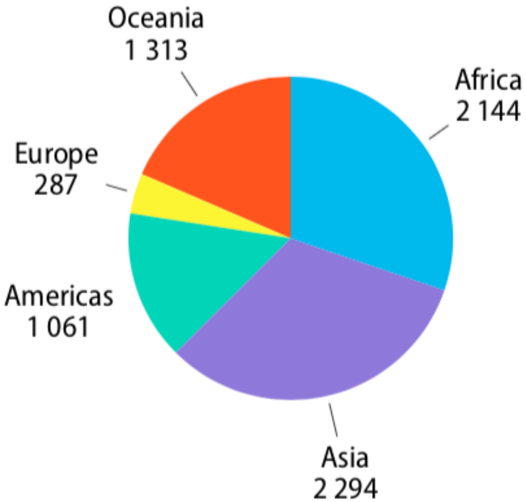
prof. Antonino Staiano

M.Sc. In ''Machine Learning e Big Data'' - University Parthenope of Naples

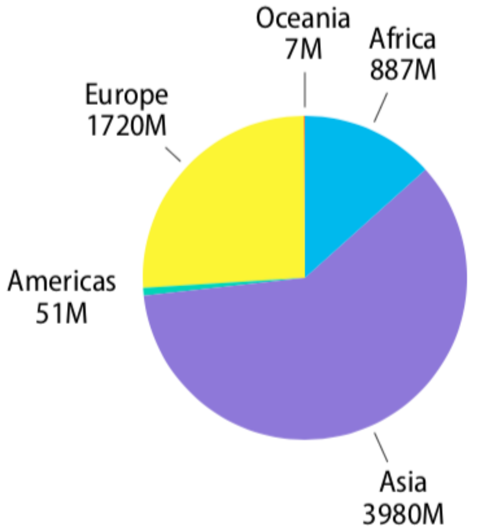
NLP tasks and their relative difficulty



Languages of the world



Languages by region of origin

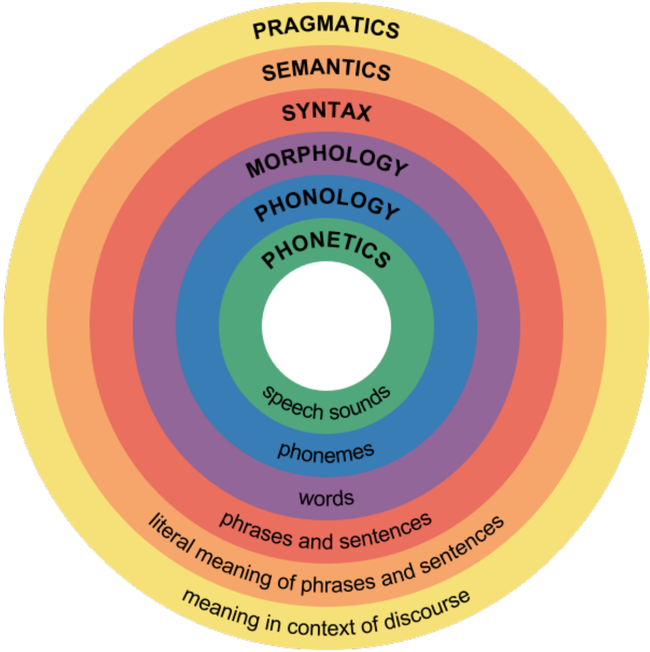


Population by region of origin

Data elaborated from Ethnologue

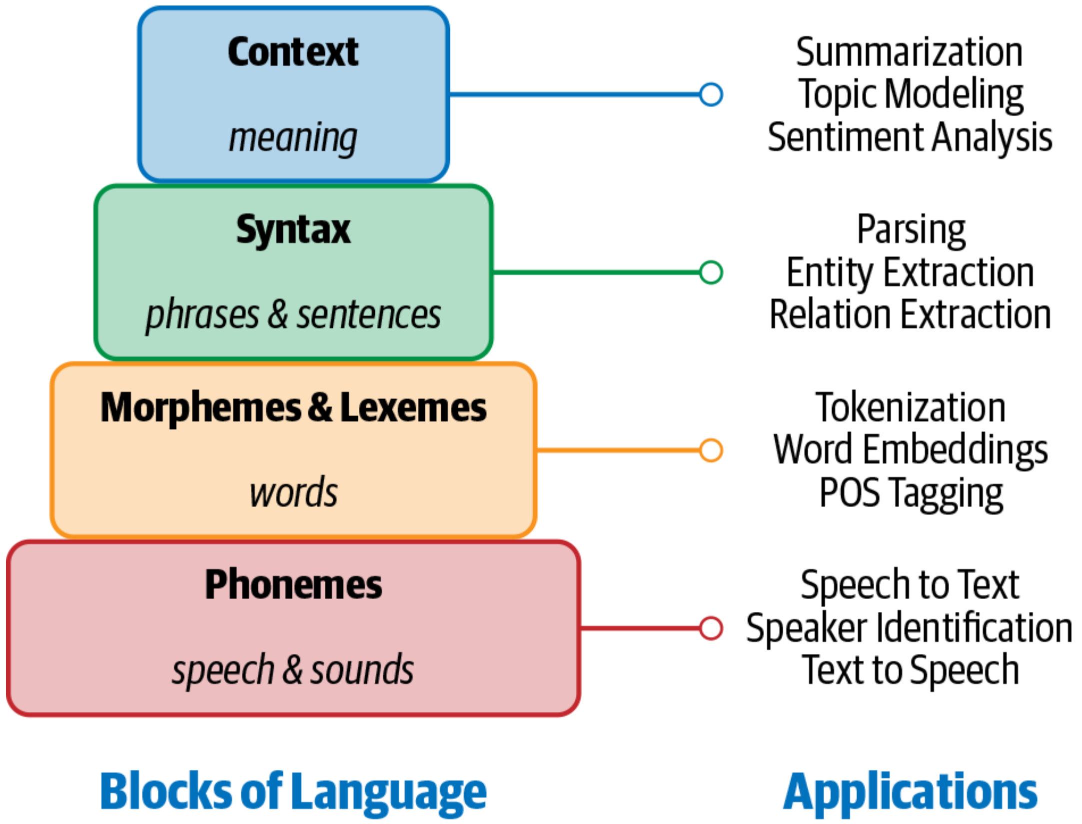
What is linguistics?

- Linguistics is the **scientific study** of language, and in particular the relationship between language form and language meaning
- Besides form and meaning, another important subject of study for linguistics is how language is **used in context**
- Noam Chomsky, sometimes called "the father of modern linguistics"
  - an American scientist who has started the development of a new framework for the study of language and is one of the founders of the field of **cognitive science**



https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

Building blocks of language and applications



Phonetics

Phonetics

- The human vocal tract can produce a wide range of sounds
  - But only certain sounds are selected as significant for communication
  - To identify and describe those sounds, we focus on each individual sound segment within a stream of speech
- The general study of the characteristics of speech sounds is called phonetics
  - Articulatory phonetics
    - How speech sounds are made or articulated
  - Acoustic phonetics
    - Physical properties of speech as sound waves
  - Auditory phonetics
    - Perception, via the ear, of speech sounds
- We exploit an already established framework for the study of speech segments known as IPA (International Phonetic Alphabet)

Consonants

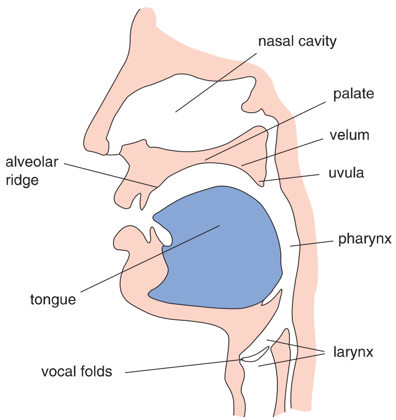
- When we describe the articulation of a consonant, the focus is on three features
  - The voice/voiceless distinction
  - The place of articulation
  - The manner of articulation

Voiced and voiceless sounds

- To make a consonant sound:
  - Air pushed out by the lungs up through the trachea to the larynx
  - Inside the larynx, the vocal cords take two basic positions
    - Vocal cords spread apart -> no obstruction for the air passing
      - **Voiceless sounds**
    - Vocal cords drawn together -> the air repeatedly pushes them apart as it passes through with a vibration effect
      - **Voiced sounds**
- To feel the distinction, try to place a fingertip gently on top of your Adam's apple and produce
  - Z-Z-Z-Z or V-V-V-V (voiced sounds)-> vibration
  - S-S-S-S or F-F-F-F (voiceless sounds) -> no vibration

Place of articulation

- After the larynx, the air enters the vocal tract via the pharynx
- It is the pushed out through the mouth and/or the nose
- Most consonant sounds are produced by using the tongue and other parts of the mouth
- The terms used to describe many sounds denote the place of articulation of the sound
  - The location inside the mouth at which the constriction take place
- To describe the place of articulation of most consonant sounds, we can start at the front of the mouth and work back
  - We consider the voiced-voiceless distinction and use the symbols of the IPA for specific sounds
  - The symbols are enclosed within [ ]



Place of articulation

- Familiar symbols
  - [p] is used for the **voiceless** consonant in *pop*, [b] in *Bob*, [m] in *mom*
  - [w] for **voiced** in *wet*
    - Bilabial consonants (made with both lips)
  - [f] and [v] are used for labiodentals, i.e., formed using upper front teeth and lower lip at the beginning of *fat* and *vat*
  - The voiceless [f] is at beginning and the voiced [v] is at the end of the pronunciation of five
  - Alveolar sounds (front of the tongue raise to **alveolar ridge**) of [t] in *tot*, [d] in *dad*, [s], [z] in *size*, [r], [l] in *rail* and [n] in *nun*
    - [t] and [s] are voiceless, [d], [z], [r], [l] and [n] are voiced
- Unfamiliar symbols
  - Think of the "th" sounds in English
    - We use [θ] for the voiceless version as in *thin* and *wrath*, and at beginning of and end of "*three teeth*"
    - We use [ð] called "eth" for the voiced version as in *thus*, *then*, *feather* and *loathe*
      - Called dentals because teeth are involved
      - If these sound are made with the tongue tip between the teeth, they are described as interdentals

Place of articulation summary

Consonants	Voiceless	Voiced	Place of articulation
Bilabials	[p]	[b], [m], [w]	both (=bi) lips (=labia)
	p**et, ta**pe	***b**et, **m**et, **w**et*	
Labiodentals	[f]	[v]	upper teeth with lower lip
	f**at, sa**fe	***v**at, sa**ve***	
Dentals	[θ]	[ð]	tongue tip behind upper teeth
	th**in, ba**th	***th**en, ba**the***	
Alveolars	[t], [s]	[d], [z], [n], [l], [r]	tongue tip to alveolar ridge
	***t**op, **s**it*	***d**og, **z**oo, **n**ut, **l**ap, **r**ap*	
Palatals	[ʃ], [ʧ]	[ʒ], [ʤ], [j]	tongue and palate
	***sh**ip, **ch**ip*	*ca**s**ual, **g**em, **y**et*	
Velars	[k]	[g], [ŋ]	back of tongue and velum
	c**at, ba**ck	***g**un, ba**ng***	
Glottals	[h]		space between vocal folds
	***h**at, **wh**o*		

Transcribing sounds

- Written English poor guide for pronunciation
- *Bang* and *tongue* end with [ŋ] ("angma") only, and there is no [g] sound (despite the spelling)
- There are some single sounds that are represented in spelling two letters
  - In *ship* we pronounce [ʃ] ("sh") no an [s] sound followed by an [h] sound
- Some similar sounds can have very different spellings
  - *Photo* and *enough*
    - Both pronounced as [f]
- There are also words with letters that are not pronounced at all
  - *Write* and *right*
    - Pronounced as [**rait**]
- Tricky letters that suggest one sound but are pronounced with another
  - *Face* vs *phase* and *race* vs *raise* ("**ce**" like [s] and "**se**" like [z])

Manner of Articulation

- With respect to the place of articulation, [t] and [s] are similar (voiceless alveolars)
- However, they are different sounds, since they differ in their manner of articulation (pronounce)
  - [t] sound is a stop consonant
    - Blocking the airflow very briefly, then letting it go abruptly
  - [s] sound is a fricative consonant
    - Pronounced by almost blocking the airflow, then letting the air escape through a narrow gap, creating friction

Consonants	Voiceless	Voiced	Manner of articulation
Stops	[p], [t], [k]	[b], [d], [g]	block airflow, let it go abruptly
	pet, talk	*bed, dog*	
Fricatives	[f], [θ], [s], [ʃ], [h]	[v], [ð], [z], [ʒ]	almost block airflow, let it escape through a narrow gap
	faith, house, she,	*vase, the, rouge*	
Affricates	[ʧ]	[ʤ]	combine a brief stop with a fricative
	cheap, rich	*jeep, rage*	
Nasals		[m], [n], [ŋ]	lower the velum, let air flow out through nose
		morning, name	
Liquids		[l], [r]	raise and curl tongue, let airflow escape round the sides
		load, light, road, write	
Glides		[w], [j]	move tongue to or from a vowel
		we, want, yes, you	

Vowels

- Vowel sounds are produced with a relatively free flow of air
  - Typically voiced
- Place of articulation
  - Front, back, high, low areas (mouth)
- Example: pronunciation of ***heat*** and ***hit***
  - *"high, front"* vowels because the sound is made with the front part of the tongue in a raised position
  - ***Hot*** is a *"low, back"* vowel

Vowel chart for English

	Front	Central	Back
High	i		u
	ɪ		ʊ
Mid	e	ə	o
	ɛ	ʌ	ɔ
Low	æ		
		a	ɑ

Front vowels

[i] *bead, beef, key, me*

[ɪ] *bid, myth, women*

[ɛ] *bed, dead, said*

[æ] *bad, laugh, wrap*

Central vowels

[ə] *above, oven, support*

[ʌ] *butt, blood, dove, tough*

Back vowels

[u] *boo, move, two, you*

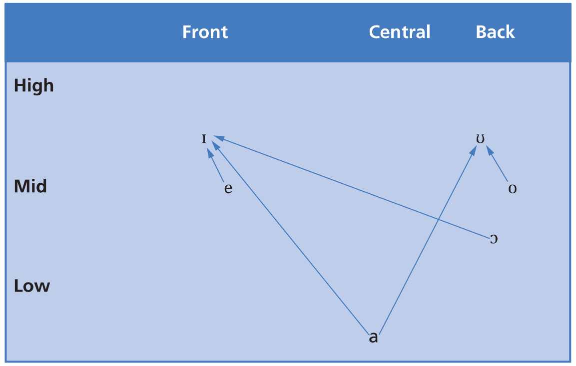
[ʊ] *book, could, put*

[ɔ] *born, caught, fall, raw*

[ɑ] *Bob, cot, swan*

Diphthongs

- Combination of two vowel sounds
- Our vocal organs move from one vocalic position [a] to another [i] as we produce the sound [ai], as in *Hi* or *Bye*
  - Movement from low to high front
- Alternatively, we can use movement from low to high back, combining [a] and [u] to produce [au]



Diphthongs

[aɪ] *buy, eye, I, my, pie, sigh*
[aʊ] *bough, doubt, cow*
[eɪ] *bait, eight, great, late, say*
[oʊ] *boat, home, owe, throw, toe*
[ɔɪ] *boy, noise, royal*

Diphthongs

- The vowels [e], [a], [o] are used
  - as single sounds in other languages and by speakers of different varieties of English
  - First sounds of diphthongs in American English
- The pronunciation of some diphthongs in *Southern British English* is different from *North American English*

	poor	*peer*	*pair*	*pour*	*pyre*	*power*
American	[pʊr]	[pir]	[peɪr]	[poʊr]	[paɪər]	[paʊər]
British	[pʊə]	[pɪə]	[pɛə]	[pɔə]	[paɪə]	[paʊə]

UNIVERSITÀ DEGLI STUDI DI NAPOLI
PARTHENOPE

Phonology

Phonology

- Since two physically different individuals would have physically different vocal tracts, every individual will pronounce sounds differently (in purely physical terms)
  - There are potentially millions of physically different ways of saying the simple word "**me**"
- Phonology is the description of the systems and patterns of speech sounds in a language
  - Concerns the underlying design, the blueprint of each sound type, which vary in different physical context

Phonology

- When we think of the [t] sounds in the words *tar*, *star*, *writer*, *butter* and *eighth* as being the same, we mean that they would be represented in the same way
- In actual speech, these [t] sounds are all potentially very different from each other because they can be pronounced in such a different ways in relation to the other sounds around them
- However, all these articulation differences in [t] sounds are less important to us than the distinction between the [t] sounds in general and the [k] sounds, or the [f] sounds, or the [b] sounds, because there are meaningful consequences related to the use of one rather than the others

Phonology

- These sounds must be distinct meaningful sounds, regardless of which individual vocal tract is being used to pronounce them, because the words *tar*, *car*, *far* and *bar are* meaningful distinct
- From this point of view, phonology is concerned with the abstract representation of sounds in our minds that enables us to recognize and interpret the meaning of words
  - based on the actual physical sounds we say and hear

Phonemes

- A phoneme describes each meaning-distinguishing sounds in a language
- The phoneme /t/ is described as a sound type, of which all the different spoken versions of [t] are tokens
- N.B.: the slash marks conventionally denote a phoneme, /t/, an abstract segment, whereas the square brackets, [t], is used for each physically produced segment
- A phoneme functions contrastively
  - /f/ and /v/ are two phonemes because they are the only basis of contrast in meaning between the words *fat* and *vat* or *fine* and *vine*
  - If we change one sound in a word and there is a change of meaning, the sound are distinct phonemes

Phonemes

- The descriptive terms we use to talk about sounds can be considered features that distinguish each phoneme from the next
  - If the feature is present, we mark it with a (+) sign and if it is not present, we use a (-) sign
- Natural classes
  - /p/ is [-voice, +bilabial, +stop]
  - /k/ is [-voice, +velar, +stop]
  - /p/ and /k/ share some features they are members of a natural class of phonemes
    - They tend to behave phonologically in similar ways
    - /v/ is [+voice, +labiodental, +fricative] and is not in the same class of /p/ and /k/
    - That's why words beginning with /pl-/ and /kl-/ are common in English, but words beginning with /vl-/ or /nl-/ are not
  - This way we describe individual phonemes but also the possible sequences of phonemes in a language

DISTINCTIVE FEATURES OF FOUR ENGLISH PHONEMES

/p/	/k/	/v/	/n/
–voice	–voice	+voice	+voice
+bilabial	+velar	+labiodental	+alveolar
+stop	+stop	+ fricative	+nasal

Phonemes

- Phone and allophones
  - A phoneme is the abstract unit or sound type, and there are many different versions of that sound type produced in an actual speech
    - The latter are called phones (phonetic units in [])
  - Each phone in a set, all versions of the same phoneme, are called allophones
- Example
  - /t/ can be pronounced in several physically different ways as phone
  - The [t] sound in *tar* is pronounced with a stronger puff of air than in *star*

Phoneme	Allophones	
	[tʰ]	(*tar*)
	[ɾ]	(*writer*)
/t/		
	[ʔ]	(*butter*)
	[t̪]	(*eighth*)

Phonemes

- Minimal pair and sets
  - Phonemic distinction in a language can be tested via pairs and sets of words
  - When two words, e.g., *fan* and *van* are identical in form except for a contrast in one phoneme occurring in the same position, the two words are described as a minimal pair
  - When group of words can be differentiated, each one from the others, by changing one phoneme they are described as a minimal set

MINIMAL PAIRS AND SETS

Minimal pairs		Minimal sets
fan – van	*bath – math*	*big – pig – rig – fig – dig – wig*
bat –beat	*math – myth*	*fat – fit – feet –fete – foot – fought*
sit –sing	*myth –Mick*	*cat –can – cap – cab – cash – cadge*

UNIVERSITÀ DEGLI STUDI DI NAPOLI
PARTHENOPE

Morphology

Morphology

- In many languages what appear to be single forms actually turn up to contain many “word-like” elements
- Example
  - In Swahili (East-Africa), *nitakupenda* is something like *I will love you*
- Is this a single word? It seems to consist of several elements that in English turn up as separate words (roughly):

ni-	*ta-*	*ku-*	*penda*
I	*will*	*love*	*you*

- Morphology studies basic forms or elements in a language

Morphology

- Morpheme
  - Words form may consist of a number of elements called morphemes
- Example
  - *talks*, *talker*, *talked*, *talking* consist of one element, *talk*, and other four elements –s, -er, -ed, -ing
    - All five elements are morphemes
- Definition: A morpheme is a minimal unit of meaning or grammatical function
  - Units of grammatical function indicate past tense or plural, for example

MORPHEMES

Minimal units of meaning	Grammatical function
re- (“again”) ***new*** (“recently made”)	***-ed*** (past tense)
tour (“travel for pleasure”) ***-ist*** (“person who”)	***-s*** (plural)

Free and bound morphemes

- Two type of morphemes
  - Free morphemes
    - Can stand by themselves as single words, e.g., new, tour
  - Bound morphemes
    - Cannot stand alone and are attached to another form, e.g., re-, -ist, -ed, -s (known as affixes)
    - All affixes (prefixes and suffixes) in English, are **bound morphemes**
  - Free morphemes can generally be identified as a set of separate English word forms such as, nouns, verbs, adjectives and adverbs
    - When they are used with bound morphemes attached, the **basic word forms** are known as stems

	undressed			***carelessness***	
un-	*dress*	*-ed*	*care*	*-less*	*-ness*
prefix	stem	suffix	stem	suffix	suffix
(bound)	(free)	(bound)	(free)	(bound)	(bound)

Lexical and functional morphemes

- Free morphemes fall in two categories
  - Lexical morphemes
    - Set of ordinary **nouns** (*girl*, *house*), **verbs** (*break*, *sit*), **adjectives** (*long*, *sad*) and **adverbs** (never, *quickly*)
    - Words that carry the content of the message we convey
    - We can add new lexical morphemes to the language , so they are an *open* class of words
  - Functional morphemes
    - **Articles** (*a*, *the*), **conjunctions** (*and*, *because*), **prepositions** (*on*, *near*) and **pronouns** (*it*, *me*)
    - We never add new functional morphemes to the language, so they are described as a closed class of words

Derivational morphemes

- The set of affixes making up the **bound morpheme class** is divided in derivational and inflectional morphemes
  - Derivational morphemes
    - Use of bound forms to make new words or to make words of a different grammatical category from the stem
      - Adding the derivational morpheme –ment changes the **verb** encourage to the **noun** encouragement
      - The noun class can become verb classify by adding the derivational morpheme –ify
      - Derivational morphemes can also be prefix, for instance, re-, pre-, ex-, mis-, co-, un-

Inflectional morphemes

- Inflectional morphemes
  - Indicate the grammatical function of a word
    - Used to show if a word is plural or singular, past tense or not, if it is a comparative or possessive form
  - English has only eight inflectional morphemes, all suffixes

*Jim'**s** two sister**s** are really different.*

*One like**s** to have fun and is always laugh**ing**.*

*The other enjoy**ed** school as a child and has always be**en** very serious.*

*One is the loud**est** person in the house and the other is quiet**er** than a mouse.*

DERIVATIONAL AND INFLECTIONAL MORPHEMES

	Nouns	Verbs	Adjectives
Derivational	*critic-**ism***	*critic-**ize***	*critic-**al***
	*encourage-**ment***	*class-**ify***	*wonder-**ful***
Inflectional	*Jim-**'s***	*like-**s**, laugh-**ing***	*quiet-**er***
	*sister-**s***	*enjoy-**ed**, be-**en***	*loud-**est***

Morphological description

- An inflectional morpheme never change the grammatical category of a word
  - *Old* and *older* are both **adjectives** (-er simply creates a different version of the adjective)
    - From Old English (-ra)
- A derivational morpheme can change the grammatical category of a word
  - *Teach* (**verb**) becomes *Teacher* (**noun**) if we add the derivational morpheme –er
    - From Old English (-ere)
  - The suffix –er in Modern English can be an inflectional morpheme (as part of an adjective) and also a distinct derivational morpheme (as part of a noun)

Morphological description

- If derivational and inflectional suffixes are used together, they always appear in that order
  - Example
    - First derivational (-er) is attached to teach, then the inflectional (-s) is added to produce teachers
- Example: "*The teacher's wildness shocked the girls' parents*"
  - We can identify thirteen morphemes

The	*teach*	*-er*	*-'s*	*wild*	*-ness*
functional	lexical	derivational	inflectional	lexical	derivational

shock	*-ed*	*the*	*girl*	*-s'*	*parent*	*-s*
lexical	inflectional	functional	lexical	inflectional	lexical	inflectional