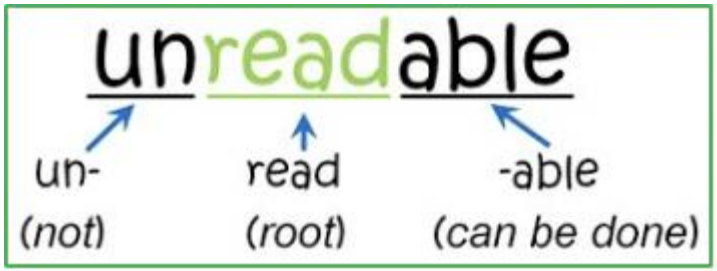
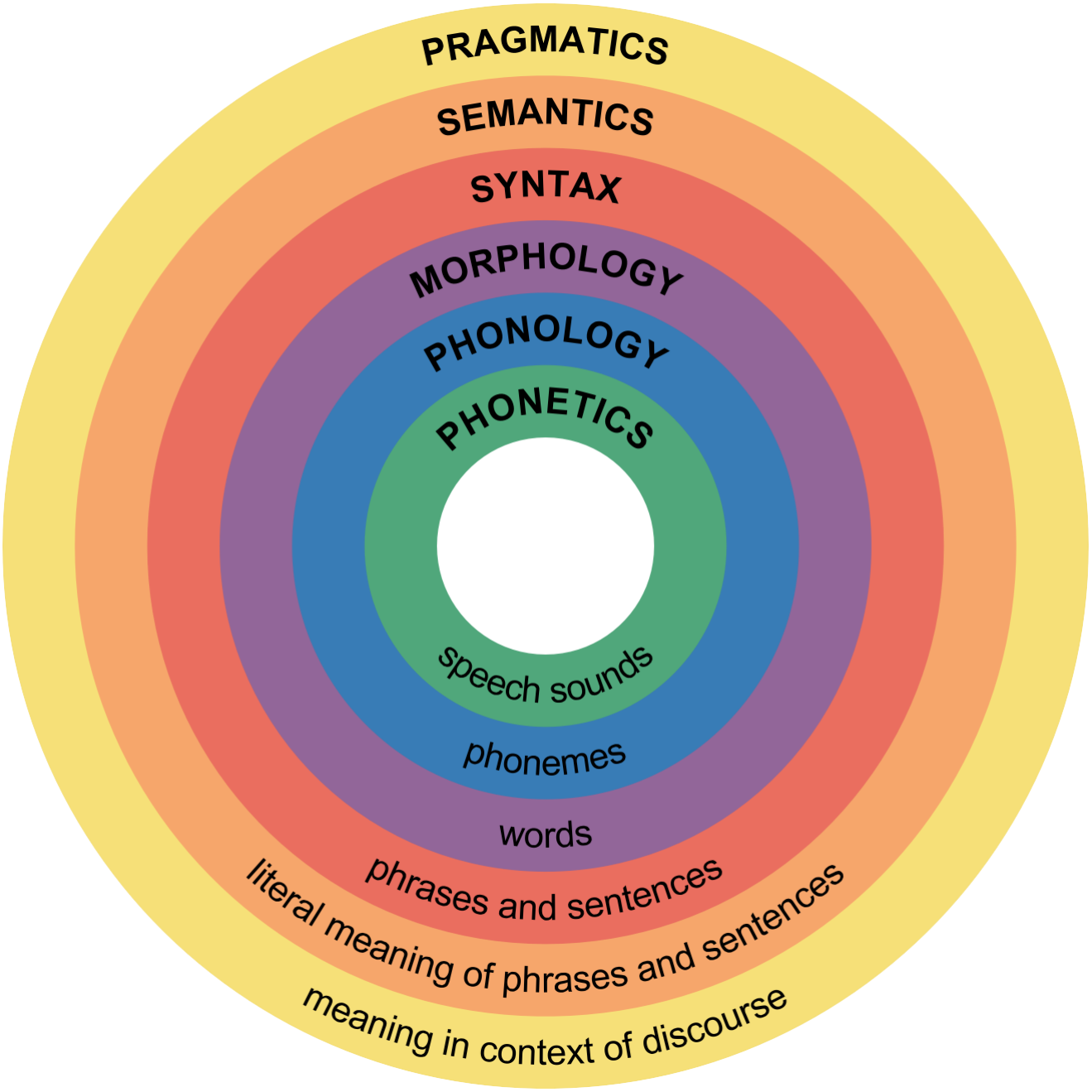
Basic Text Processing: Words & Morphology

(Natural Language Processing)

H.N.D. Thilini
hnd@ucsc.cmb.ac.lk



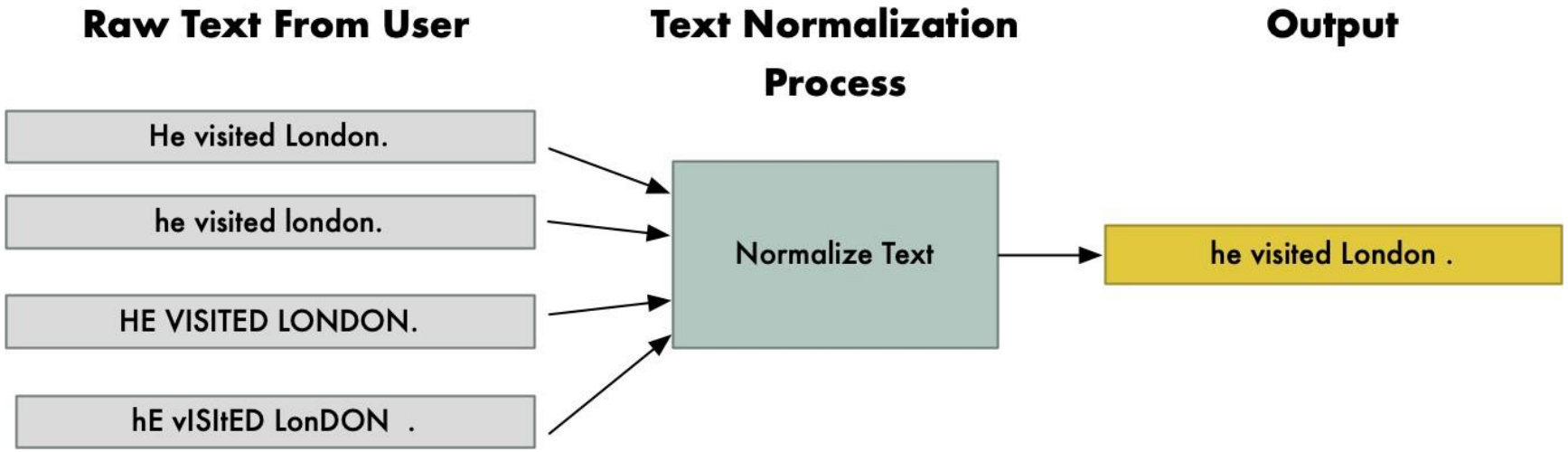

PRAGMATICS
SEMANTICS
SYNTAX
MORPHOLOGY
PHONOLOGY
PHONETICS
speech sounds
phonemes
words
phrases and sentences
literal meaning of phrases and sentences
meaning in context of discourse

Basic Text Processing

- Every NLP task needs to do text normalization to determine what are the words of the document:
  - Segmenting/tokenizing words in running text
    - Special characters like hyphen "-" and apostrophe ‘
  - Normalizing word formats
    - (Non) capitalization of words
    - Reducing words to stems or lemmas
  - Segmenting sentences in running text
- To do these tasks, we need to use morphology

Text Normalization

- Text normalization reduces variations in word forms to a common form when the variations mean the same thing.

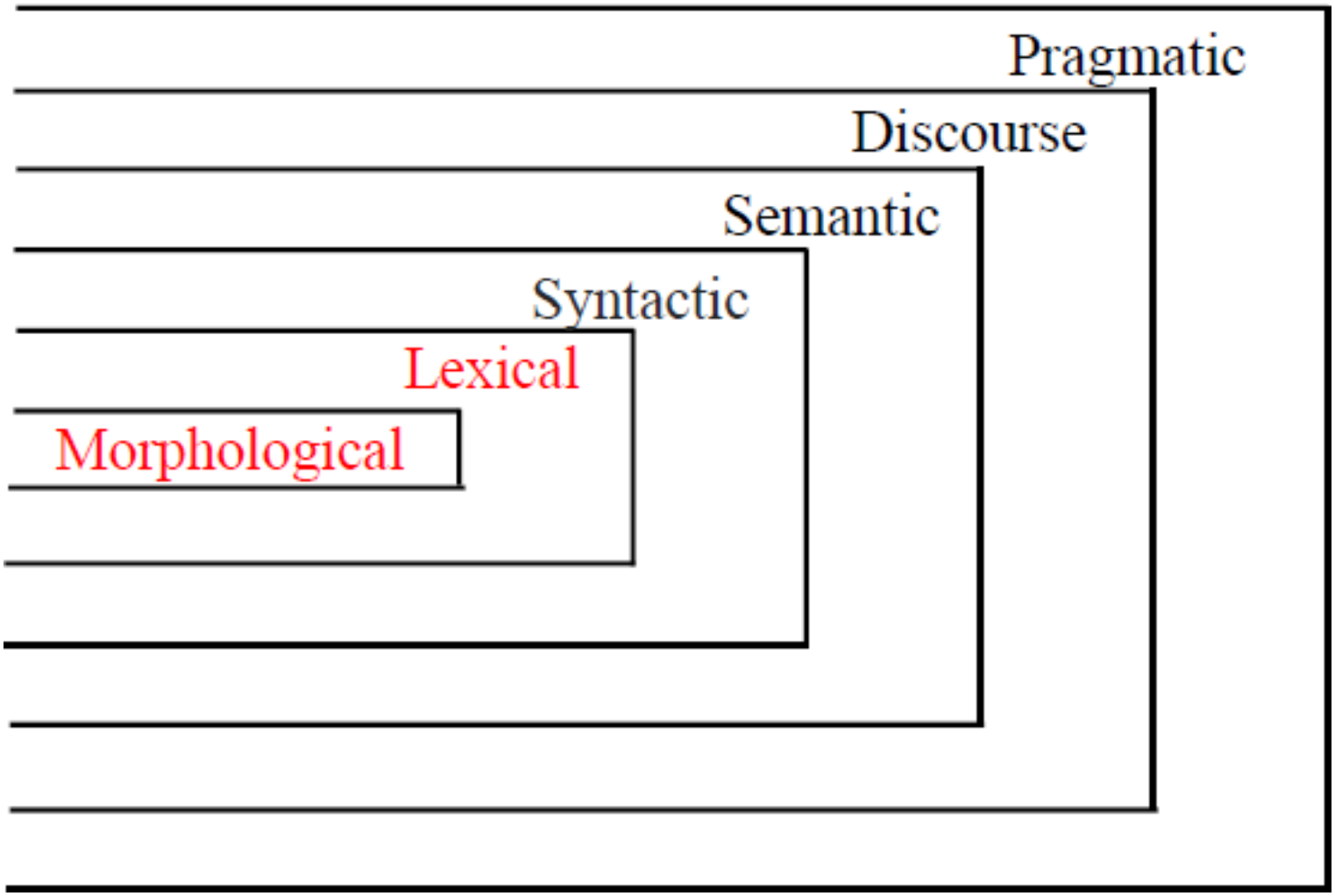


Text Normalization

- Text normalization reduces variations in word forms to a common form when the variations mean the same thing.

- Stemming and Lemmatization:
  - Stemming just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling.
  - Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma.

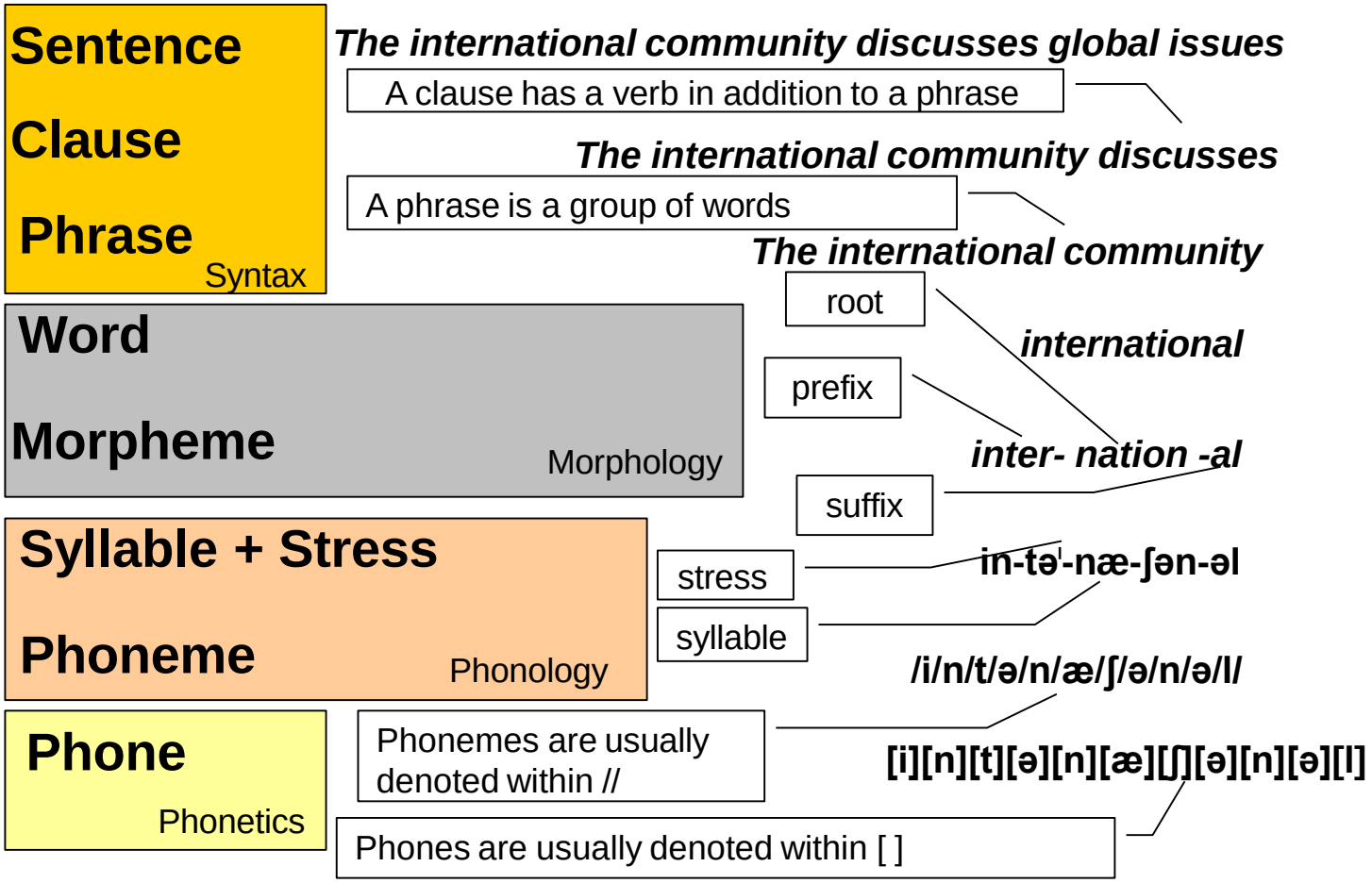
Synchronic Model of Language



Word!

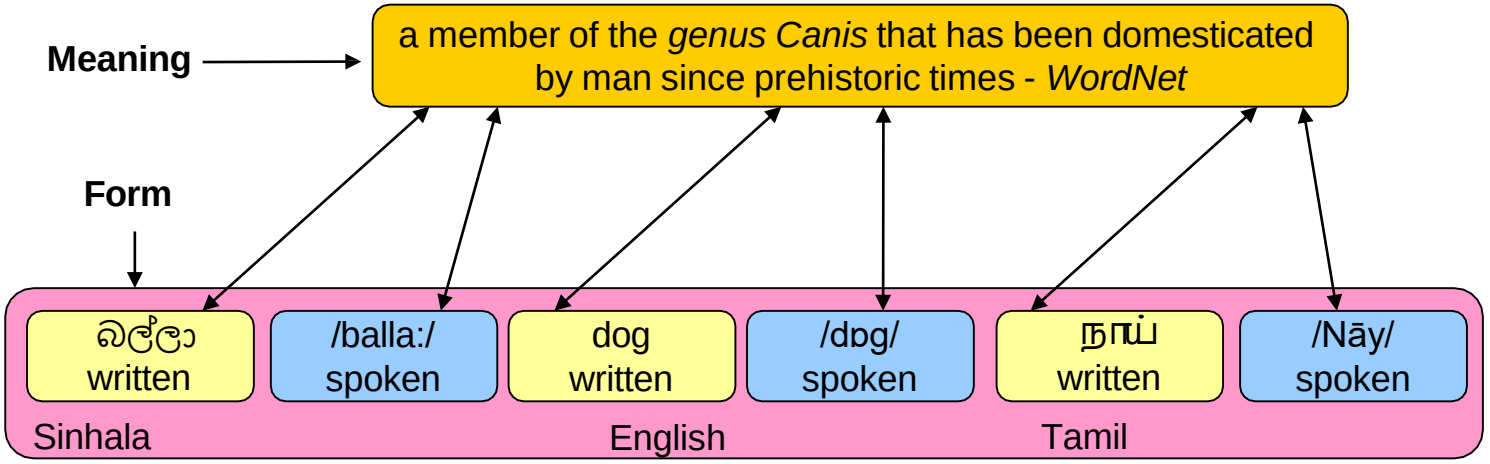
- In formal languages, words are arbitrary strings
- In natural languages, words are made of meaningful sub units called morphemes
- How we define a word?

Word in the Context of Linguistic Analysis



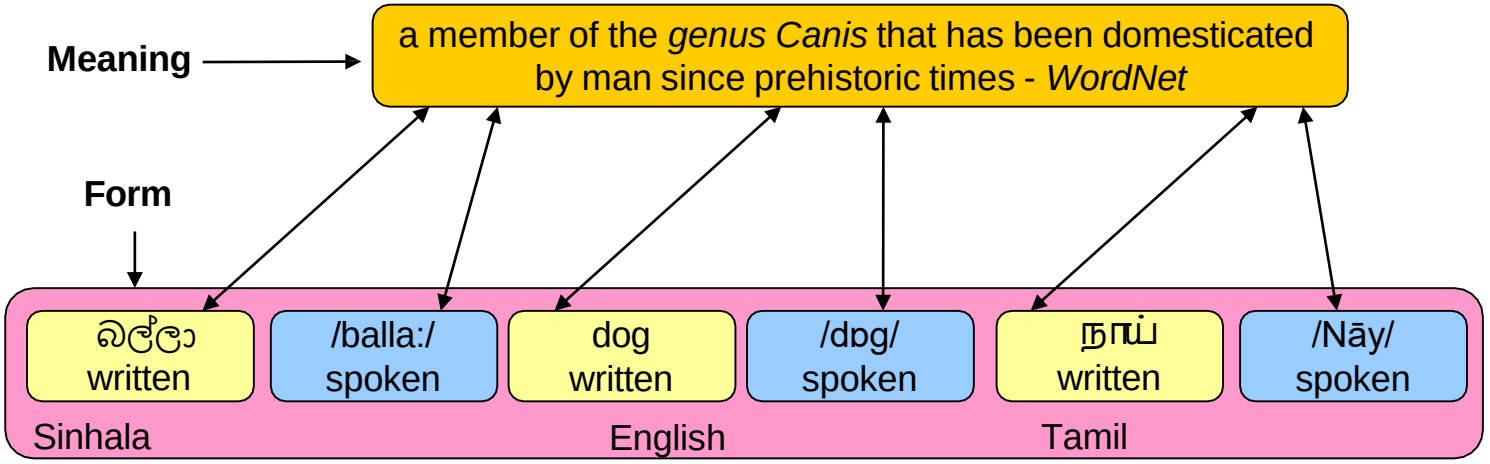
What is a Word?

- A word is a unique linguistic entity, which can be expressed in either speech or writing - *word has a* ***form***
  - ***Spoken* vs. *Written***
- A word is a linguistic device that relates ***form*** and ***meaning***
- The relation between *form* and *meaning* is arbitrary
  - The concept DOG has arbitrary spoken and written forms in different languages



What is a Word?

- A word is a unique linguistic entity, which can be expressed in either speech or writing - *word has a* ***form***
  - ***Spoken* vs. *Written***
- A word is a linguistic device that relates ***form*** and ***meaning***
- The relation between *form* and *meaning* is arbitrary
  - The concept DOG has arbitrary spoken and written forms in different languages



Lexical Semantics

- The ***meaning*** of a word is largely determined by its context
- Consider these two uses of the word *bank*
  - Instead, a *bank* can hold the investments in a custodial account in the client's name
  - But as agriculture burgeons on the east *bank*, the river will shrink even more
- These two contextual variations of the meaning *bank* are called (***word***) ***senses***
- Therefore, a sense is an aspect of the meaning of a word, usually denoted by ***bank***^{1} and ***bank***^{2}

Relations between Senses

- When the meaning of two senses of two different words are identical or nearly identical they are called **synonyms**.
  - couch/sofa, beautiful/pretty, car/automobile
- Word with opposite meanings are **antonyms**
  - long/short, big/little, fast/slow, cold/hot, dark/light, rise/fall, up/down
- One sense is a **hyponym** of another sense if the first sense is more specific, denoting a subclass of another
  - car is a hyponym of vehicle, dog is a s of animal
- One sense is a **hypernym** of another sense if the first sense is more general, denoting a subclass of another
  - vehicle is a hypernym of car , animal is a hypernym of dog

Relations between Senses

- One sense is a **meronym** of another sense if the first sense is part of the second
  - leg is a *meronym* of chair, wheel is a meronym of of car
- One sense is a **holonym** of another sense if the first sense has *a part of* denoted by the second,
  - Chair is a *holonym* of leg, car is a *holonym* of wheel

Comparison of *Meaning*, *Written* and *Spoken Forms*

Meaning	Written Form	Spoken Form	Name	Examples
Different	Different	Different	Different	cat,dog
Different	Different	Same	Homophones	bear , bare
Different	Same	Different	Homographs	bass- fish, bass- music
Different	Same	Same	Homonyms	bank
Same	Different	Different	Synonyms	high, tall
Same	Different	Same	Orthographic Variants	labor, labour
Same	Same	Different	Phonetic Variants	either /iy dh er/ , /ay dh er/
Same	Same	Same	Identical	-

Exercise: Identify Sinhala/Tamil examples related to the above table

Morphology

- **Morphology** is the field of linguistics that studies the internal structure of words
- Study of the rules that govern the combination of morphemes.
- How words are built up from smaller meaningful units called morphemes
- morph = shape, Logy = logos = study of

Morphology

- General morphological theory applies to all languages as all natural human languages have systematic ways of structuring words (even sign language)

- Must be distinguished from morphology of a specific language
  - English words are structured differently from German words, although both languages are historically related
  - Both are vastly different from Arabic

Morphemes

- Smallest units of meaning
- Express concepts or relationships
  - Ex: car, table, anti-, re-.
- Express syntactic features
  - number (singular, plural)
  - tense (present, past, future)
  - gender (masculine, feminine)
  - case (nominative, accusative, genitive, dative, locative, ablative, instrumental, vocative) .

Morphemes

- Morph:
  - Morphemes as parts of a word.
    - Car – the morpheme *car* is realized as the morph *car* to form the word *car*.
    - Cars – the morpheme *car* and the plural morpheme are realized as *car* and +*s* respectively, to form the word *cars.*
- Allomorphs:
  - The different forms of a morpheme.
    - Ex: the plural morpheme in English has several allomorphs (+es, +s, stem vowel alteration, etc.).
    - Ex: take, took.

Morphemes: Types

- Free morphemes
  - Can form words by themselves.
    - Ex: Car, dog.
- Bound morphemes
  - Must be combined with other morphemes to form words.
    - Ex: Plural morpheme, anti-.
- Words can be formed by free morphemes only, or free and bound morphemes.

Morphemes: 2 classes

- We can usefully divide morphemes into two classes
  - Stems: The core meaning bearing units
  - Affixes: Bits and pieces that adhere to stems to change their meanings and grammatical functions

Affixes

- **Prefixes** appear in front of the stem to which they attach
  - **un**- + *happy* = ***un****happy*
- **Infixes** appear inside the stem to which they attach
  - **blooming**- + *absolutely* = *abso****blooming****lutely*
- **Suffixes** appear at the end of the stem to which they attach
  - *Happy* + -**ness** = *Happiness*
  - English may stack up to 4 or 5 suffixes to a word
  - Agglutinative languages like Turkish may have up to 10
- **Circumfixes** appear at both the beginning and end of stem
  - German past participle of *sagen* is *gesagt*: **ge**- + *sag* + -**t**
- Spelling and sound changes often occur at the boundary
  - Very important for NLP

Two Broad Classes of Morphology

- Inflectional Morphology
- Derivational Morphology

Inflection

- Inflection modifies a word's form in order to mark the grammatical subclass to which it belongs
  - *apple* (singular) > *apple**s*** (plural)

- Inflection does not change the grammatical category (part of speech)
  - *apple* – noun; *apples* – still a noun

- Inflection does not change the overall meaning
  - both *apple* and *apples* refer to the fruit

Think examples in your own language!

Derivation

- Derivation creates a new word by changing the category and/or meaning of the base to which it applies
  - *create* + -**tion** = *crea**tion***
- Derivation can change the grammatical category (part of speech)
  - *sing* (verb) =/= *sing**er*** (noun)
- Derivation can change the meaning
  - *judge* (form an opinion) =/= *judg**ment*** (ability to make considered decisions)
- Derivation is often limited to a certain group of words
  - You can **Clintonize** the government, but you can't **Bushize** the Government
  - 'mahindakaranaya'?
  - This restriction is partially phonological

Inflection & Derivation: Order

- **Order is important** when it comes to inflections and derivations
  - **Derivational suffixes must precede inflectional suffixes**
    - *sing* + *-er* + *-s* is ok
    - *sing* + *-s* + *-er* is not
  - This order may be used as a clue when working with natural language text

Inflection & Derivation in English

- English has few inflections
  - Many other languages use inflections to indicate the role of a word in the sentence
  - Use of case endings allows fairly free word order
  - English instead has a fixed word order
    - Position in the sentence indicates the role of a word, so case endings are not necessary
  - This was not always true; Old English had many inflections

- English has many derivational affixes, and they are regularly used to form new words
  - Part of this is cultural -- English speakers readily accept newly introduced terms

Inflection & Derivation in English

- examples from J&M, sections 3.1 – 3.3 (2nd ed.)

Morphological Form Classes	Regularly In ected Verbs			
stem	walk	merge	try	map
-s form	walks	merges	tries	maps
-ing participle	walking	merging	trying	mapping
Past form or *-ed* participle	walked	merged	tried	mapped

Inflection & Derivation in English

A very common kind of derivation in English is the formation of new nouns, often from verbs or adjectives. This process is called **nominalization**. For example, the suffix *-ation* produces nouns from verbs ending often in the suffix *-ize* (*computerize* → *computerization*). Here are examples of some particularly productive English nominalizing suffixes.

Suffix	Base Verb/Adjective	Derived Noun
-ation	computerize (V)	computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

Adjectives can also be derived from nouns and verbs. Here are examples of a few suffixes deriving adjectives from nouns or verbs.

Suffix	Base Noun/Verb	Derived Adjective
-al	computation (N)	computational
-able	embrace (V)	embraceable
-less	clue (N)	clueless

- examples from J&M, sections 3.1 – 3.3 (2nd ed.)

Classes of Words

- **Closed** classes are fixed – new words cannot be added
  - Pronouns, prepositions, comparatives, conjunctions, determiners (articles and demonstratives)
  - Function words
- **Open** classes are not fixed – new words can be added
  - Nouns, Verbs, Adjectives, Adverbs
  - Content words
  - New content words are a constant issue for NLP

Creation of New Words

- **Derivation** - adding prefixes or suffixes to form a new word
  - *Clinton -> Clintonize*
- **Compounding** - combining two existing words
  - *home + page ->homepage*
- **Clipping** - shortening a polysyllabic word
  - *Internet -> net*
  - *Examination ➔exam*
- **Acronyms** - take initial sounds or letters to form new word
  - *Unesco -> United Nations Educational, Scientific and Cultural Organization*
- **Blending** - combine parts of two words
  - *motor + hotel -> motel*
  - *smoke + fog -> smog*
- **Backformation**
  - *resurrection -> resurrect*
  - *Editor -> Edit*

Word Formation Rules: Agreement

- Plurals
  - In English, the morpheme s is often used to indicate plurals in nouns
  - Nouns and verbs must agree in plurality

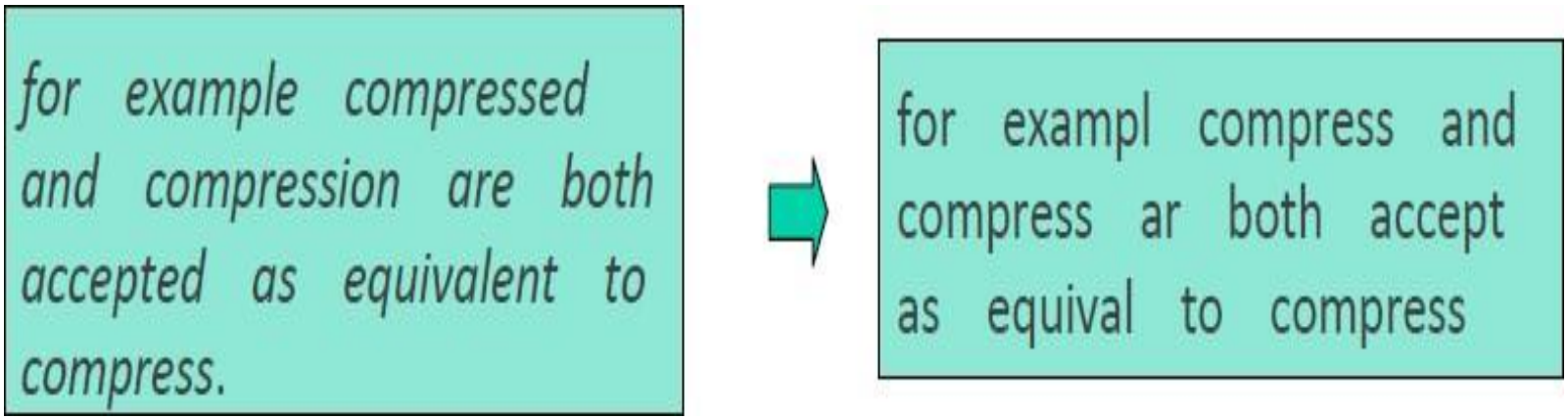
- Gender – nouns, adjectives and sometimes verbs in many languages are marked for gender
  - 2 genders (masculine and feminine) in Romance languages like French, Spanish, Italian
  - 3 genders (masc, fem, and neuter) in Germanic and Slavic languages
  - More are called noun classes – Bantu has up to 20 genders
  - Gender is sometimes explicitly marked on the word as a morpheme, but sometimes is just a property of the word

How does NLP make use of morphology?

- Stemming
  - Strip prefixes and / or suffixes to find the base root, which may or may not be an actual word
    - Spelling corrections are not made
- Lemmatization
  - Strip prefixes and / or suffixes to find the base root, which will always be an actual word
    - Spelling corrections are crucial
    - Often based on a word list, such as that available at WordNet
- Morphological parsing
  - Knowledge of morphemes for a particular language can be a powerful aid in guessing the part of speech and grammatical features for even unknown terms

Stemming

- Removal of affixes (usually suffixes) to arrive at a base form that may or may not necessarily constitute an actual word
- Continuum from very conservative to very liberal modes of stemming
  - Very Conservative
    - Remove only plural –s
  - Very Liberal
    - Remove all recognized prefixes and suffixes



Porter Stemmer

- Popular stemmer based on work done by Martin Porter
  - *M.F. Porter. An algorithm for suffix stripping. 1980, Program 14(3), pp. 130-137*.
- Very liberal step stemmer with five steps applied in sequence
  - See example rules on next slide
- Probably the most widely used stemmer
  - Has been incorporated into a number of Information Retrieval systems
- Does not require a lexicon.
- Open source software available for almost all programming languages.

Rules of Porter Stemmer

Step 1a

sses	→ ss !	caresses	→ caress!	
ies	→ i !	ponies	→ poni!	
ss	→ ss !	caress	→ caress!	
s	→ ø	cats	→ cat!	
		walking	→ walk!	
		sing	→ sing!	
…!				

Where *v* is the occurrence of any vowel.

From Dan Jurafsky

Step 2 (for long stems)

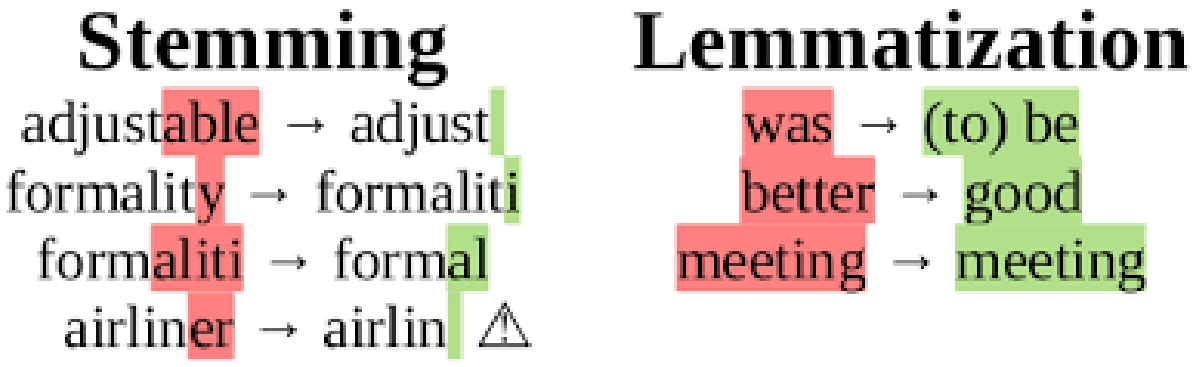
ational→	ate	relational	→ relate!
izer→	ize !	digitizer	→ digitize!
ator→	ate !	operator	→ operate!
…!			

al	→ ø	revival	→ reviv!
able	→ ø	adjustable	→ adjust!
ate	→ ø	activate	→ activ!
…!			

https://www.youtube.com/watch?v=Vx72Q5Jqc5M

Lemmatization

- Removal of affixes (typically suffixes),
- But the goal is to find a base form that does constitute an actual word
- Example:
  - *parties* -> remove *-es*, correct spelling of remaining form -> *party*
- Spelling corrections are often rule-based
- May use a lexicon to find actual words



Morphological Parsing

- English is continuously gaining new words on a daily basis
- And new words are a problem for many NLP systems
  - New words won't be found in the MRD or lexicon, if one is used
- How might morphology be used to help solve this problem?
- What part of speech are:
  - clemness
  - foramtion
  - depickleated
  - outtakeable

MRD: Machine Readable Dictionary

Problem of Ambiguous Affixes

- Some affixes are ambiguous:
  - **-er**
    - Derivational: Agentive –er Verb + -er > noun (*sing* + *-er* -> *singer*)
    - Inflectional: Comparative –er Adjective + -er > Adjective (*big* + *-er* -> *bigger*)
  - **-s** or **–es**
    - Inflectional: Plural Noun + -(e)s > Noun (*cat* + *-s* -> *cats*)
    - Inflectional: 3rd person sing. Verb + -(e)s > Verb (*write* + *-s* -> *writes*)
  - **-ing**
    - Inflectional: Progressive Verb + -ing > Verb *he is swimming*
    - Derivational: "act of" Verb + -ing > Noun *swimming is good for health*
    - Derivational: "in process of" Verb + -ing > Adjective *swimming pool*

This morphological ambiguity creates a problem for NLP

Complex Morphology

- Some languages requires complex morpheme segmentation
  - In Turkish,

  - Uygarlastiramadiklarimizdanmissinizcasina
  - `(behaving) as if you are among those whom we could not civilize'
  - Uygar `civilized' + las `become'
    + tir `cause' + ama `not able'
    + dik `past' + lar 'plural'
    + imiz 'p1pl' + dan 'abl'
    + mis 'past' + siniz '2pl' + casina 'as if'

Computational Morphology

- Analysis (words → encoded meaning)
  - Take a sequence of characters as input, and produce an analysis of the information encoded in the characters.
    - Ex: *Plays* -> (play/noun/plural) or (play/verb/3rd person/singular/present)
- Generation (meaning → words)
  - Generate words from a set of features.
    - Ex: (run/verb/1st person/singular/past) -> *ran*

What is a Corpus?

- Corpus is a collection of written texts, especially the entire works of a
  particular author or a body of writing on a particular subject.
- In linguistics, a corpus or text corpus is a large and structured set of texts.
  - Google Books Ngram Corpus
  - American National Corpus - 22 million words of written and spoken data
  - COBUILD corpus - 4.5 billion words
  - British National Corpus - 100-million-word
  - Corpus of Contemporary American English (COCA) - 425 million words
  - UCSC 10M words Sinhala Corpus – 10 million words

What is Corpus Linguistics?

- A methodology to process text and provide information about the text, usually at the one or two word level.

- Statistical analysis
  - Word frequencies
  - Collocations
  - Concordances

Preliminary Text Processing Required:

- Find the words:
  - Filter out ‘junk data’
    - Formatting / extraneous material
    - First be sure it doesn’t reveal important information
  - Deal with upper / lower case issues
  - Tokenize
    - Decide how you define a ‘word’
    - How to recognize and deal with punctuation
      - Apostrophes (one word it’s vs. two words it’s)
      - Hyphens (snow-laden vs. New York-New Jersey)
      - Periods (kept with abbreviations vs. separated as sentence markers)

Preliminary Text Processing Required:

- Word segmentation
  - No white space in Japanese language
  - Compound words –
    - “Lebensversicherungsgesellschaftsangestellter”
- Additional issues if OCRed data or speech transcripts
- Morphology (To stem or not to stem?)
  - Depends on the application

Word Counting in Corpora

- After corpus preparation, additional decisions
  - Ignore capitalization at beginning of sentence? Is "They" the same word as "they"?
  - Ignore other capitalization? is "Company" the same word as "company"
  - Stemming? Is "cat" the same word as "cats"
- Terminology for word occurrences:
  - Tokens – the total number of words
  - Distinct Tokens (sometimes called word types) – the number of distinct words, not counting repetitions
  - The following sentence from the Brown corpus has 16 tokens and 14 distinct tokens:

    They picnicked by the pool, then lay back on the grass and looked at the stars.

Word Frequencies

- Count the number of each token appearing in the corpus (or sometimes single document)
- A frequency distribution is a list of all tokens with their frequency, usually sorted in the order of decreasing frequency
  - Used to make “word clouds”
  - For example, http://www.tumblr.com/tagged/word+cloud
  - Create your own word cloud? https://www.wordclouds.com/
- Used for comparison and characterization of text

Word Cloud for Sinhala Songs!

How many words in a corpus?

- Let N be the number of tokens
- Let V be the size of the vocabulary (the number of distinct tokens)
- Church and Gale (1990) suggest that the ***V*** grows with at least the square root of the ***N*** (i.e. $|V| > O(N^{½})$).

	Tokens=N	Types=|V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
GoogleN-grams	1 trillion	13 million

Zipf's Law

- **Rank (r):** The numerical position of a word in a list sorted by decreasing frequency (f).
- Zipf (1949) "discovered" that:

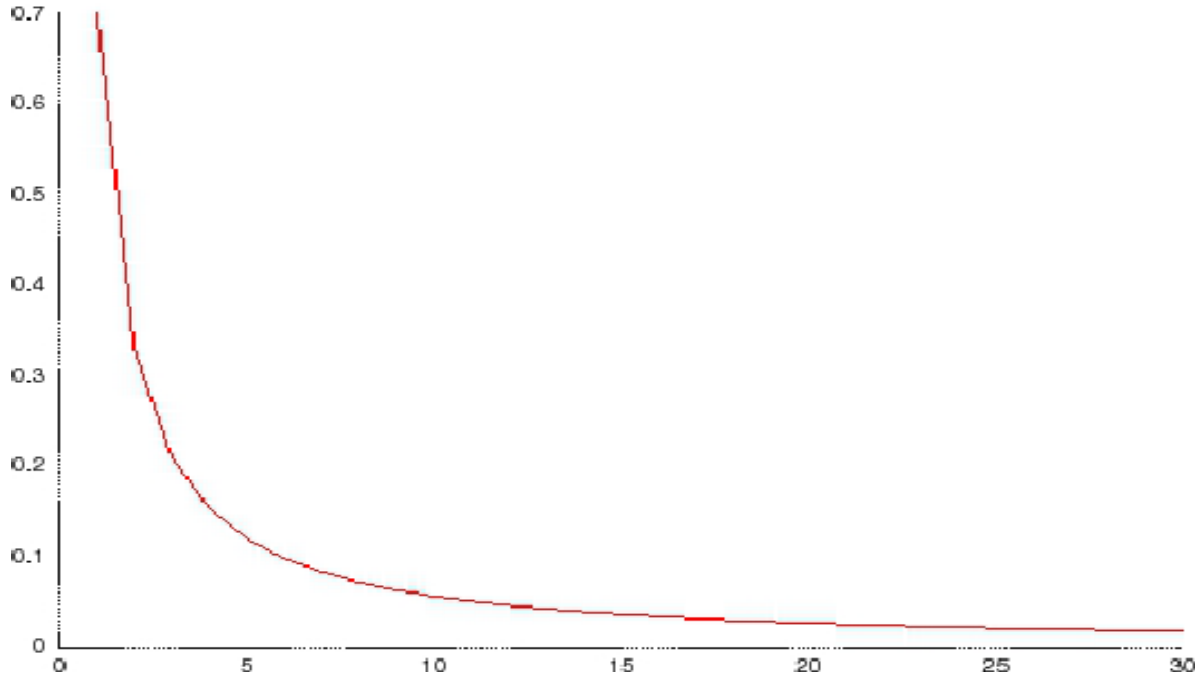
$$f \cdot r = k \text{ (for constant k)}$$

- If probability of word of rank *r* is *pr* and N is the total number of word occurrences:

$$pr = f/N = A/r \text{ for corpus independent constant } A = 0.1$$

Zipf Curve

- A typical Zipf-law rank distribution. The y-axis represents word occurrence frequency
  - x-axis represents rank (highest at the left)
  - y-axis represents the frequency of words



Zipf's Law Impact on Language Analysis

- Good News: Stop words (commonly occurring words such as "the") will account for a large fraction of text so eliminating them greatly reduces size of vocabulary in a text

- Bad News: For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.

Corpora as a Learning Tool

- Word Concordance
  - https://www.lextutor.ca/conc/eng/