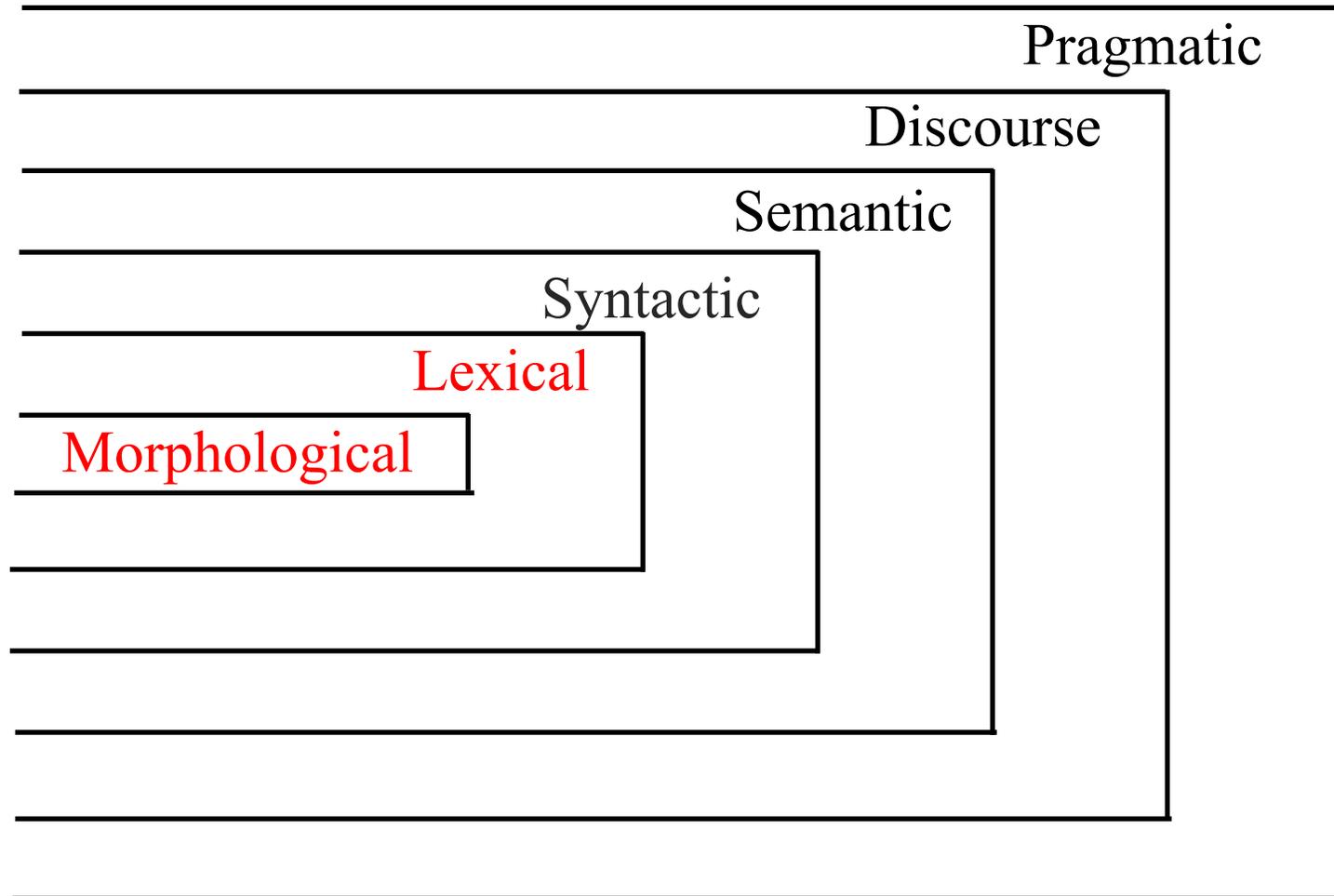

Morphology

Synchronic Model of Language



Morphology

- Morphology is the level of language that deals with the internal structure of words
- 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
- Linguists care most about general morphological theory
- NLP researchers care most about morphology of specific languages

Minimal Units of Meaning

- **Morpheme** = the minimal unit of meaning in a word
 - *walk*
 - *-ed*
- **Simple words** cannot be broken down into smaller units of meaning
 - Monomorphemes
 - Called base words, roots or **stems**
- **Affixes** are attached to free or bound forms
 - prefixes, infixes, suffixes, circumfixes

Free vs. Bound

- Units of meaning that can stand on their own are “free”
 - Words (*apple, happy*)
- Units of meaning that cannot stand on their own are “bound”
 - Prefixes (un- in *unhappy*)
 - Suffixes (-s in *apples*)
 - Contractions and clitics (in *don't*, the *t* is the clitic)

Affixes

- **Prefixes** appear in front of the stem to which they attach
 - **un-** + happy = *unhappy*
- **Infixes** appear inside the stem to which they attach
 - **-blooming-** + absolutely = *absobloominglutely*
- **Suffixes** appear at the end of the stem to which they attach
 - *emotion* = emote + -ion
 - 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

Inflection & Derivation

- **Inflectional** morphology - the way in which words vary (or ‘inflect’) in order to express grammatical contrasts in sentences, such as singular / plural and present / past tense.
 - **Inflectional affixes**
- **Derivational** morphology - the principles governing the construction of new words, without reference to the specific grammatical role a word might play in a sentence.
 - **Derivational affixes**

Inflection

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

Derivation

- Derivation creates a new word by changing the category and/or meaning of the base to which it applies
- Derivation can change the grammatical category (part of speech)
 - sing (verb) > singer (noun)
- Derivation can change the meaning
 - act of singing > one who sings
- Derivation is often limited to a certain group of words
 - You can **Clintonize** the government, but you can't **Bushize** the government
 - 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
- Look at 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
- **Acronyms** - take initial sounds or letters to form new word
 - Scuba → Self Contained Underwater Breathing Apparatus
- **Blending** - combine parts of two words
 - motor + hotel → motel
 - smoke + fog → smog
- **Backformation**
 - resurrection → resurrect

Word Formation Rules

- Word formation rules are very important for NLP
- Word formation rules apply to stems
- Stems in this sense are any form to which an affix can attach, whether simple or complex
 - Prefix + stem = new word
 - un- + adjective = adjective
 - Stem + suffix = new word
 - verb + -er = noun (agentive noun)
- Order of affixation is important

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
- Part of speech guessing
 - Knowledge of morphemes for a particular language can be a powerful aid in guessing the part of speech for an unknown term

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
 - Many points in between the two extremes
- Good resource:
 - <http://www.comp.lancs.ac.uk/computing/research/stemming/>

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
 - Contains rules like
 - ATIONAL -> ATE (e.g. relational -> relate)
 - ING -> ϵ (e.g. if the stem is a verb, motoring -> motor)
 - SSES -> SS (e.g. grasses -> grass)
- Probably the most widely used stemmer
 - Has been incorporated into a number of Information Retrieval systems
- Does not require a lexicon (unlike Finite State techniques)

Some other Stemmers for English

- Paice-Husk Stemmer
 - Simple iterative stemmer; rather heavy when used with standard rule set
- Krovetz Stemmer
 - Light stemmer; removes inflections only; removal of inflections is very accurate
 - Often used as a first step before using another stemmer for increased compression
- Lovins Stemmer
 - Single-pass, context-sensitive, longest match stemmer; not widely used
- Dawson Stemmer
 - Complex linguistically targeted stemmer based on Lovins; not widely used

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
- Krovetz “stemmer” is really a lemmatizer

Stemming vs. Lemmatization

- Does it matter whether the resulting base form is an actual word or not?
 - The answer is task-dependent
- Task:
 - Dictionary look-up: Lemmatization would be better
 - **Information retrieval**: Stemming would probably be better
 - Clustering: Stemming would probably be the better choice

Guessing the Part of Speech

- 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

Ambiguous Affixes

- Some affixes are ambiguous:
 - -er
 - Derivational: Agentive –er Verb + -er > Noun
 - Inflectional: Comparative –er Adjective + -er > Adjective
 - -s or –es
 - Inflectional: Plural Noun + -(e)s > Noun
 - Inflectional: 3rd person sing. Verb + -(e)s > Verb
 - -ing
 - Inflectional Progressive Verb + -ing > Verb
 - Derivational “act of” Verb + -ing > Noun
 - Derivational “in process of” Verb + -ing > Adjective
- As with all other ambiguity in language, this morphological ambiguity creates a problem for NLP