CSC 594 Topics in AI –
Applied Natural Language Processing
Fall 2009/2010

2. Some Basic Concepts:
   Morphology

Levels of Language Analysis
1. Phonology
   • study of sound systems of languages
2. Morphology
   • study of structure of words: the structure of words in a language, including patterns of inflections and derivations
3. Syntax
   • study of organization of words in sentences: the ordering of and relationship between the words in phrases and sentences
4. Semantics
   • study of meaning in language: the study of how meaning in language is created
5. Pragmatics
   • study of language in use: the branch of linguistics that studies language use rather than language structure
6. Discourse
   • study of language, especially the type of language used in a particular context or subject
7. World Knowledge

Some English phonemes

<table>
<thead>
<tr>
<th>English</th>
<th>IPA</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>p</td>
<td>pen, cup, lip</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>box, rib</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>test, bit</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
<td>dot, odd</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
<td>kite, ask</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
<td>get, long</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>faint, tall</td>
</tr>
<tr>
<td>s</td>
<td>s</td>
<td>sink, wax</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>izzly, wax</td>
</tr>
<tr>
<td>sh</td>
<td>ʃ</td>
<td>this, wash</td>
</tr>
<tr>
<td>zh</td>
<td>ʒ</td>
<td>this, wash</td>
</tr>
<tr>
<td>th</td>
<td>θ</td>
<td>thin, path</td>
</tr>
<tr>
<td>th</td>
<td>ð</td>
<td>thin, path</td>
</tr>
<tr>
<td>j</td>
<td>ʤ</td>
<td>jazz, open</td>
</tr>
<tr>
<td>dž</td>
<td>ʤ</td>
<td>jazz, open</td>
</tr>
<tr>
<td>r</td>
<td>r</td>
<td>run, rug</td>
</tr>
<tr>
<td>l</td>
<td>l</td>
<td>look, comb</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>nose, pen</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
<td>men, pen</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>nose, pen</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>in, end</td>
</tr>
<tr>
<td>e</td>
<td>ɛ</td>
<td>in, end</td>
</tr>
<tr>
<td>i</td>
<td>i</td>
<td>in, end</td>
</tr>
<tr>
<td>o</td>
<td>ɔ</td>
<td>in, end</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td>in, end</td>
</tr>
</tbody>
</table>

Source: IPA chart for English
2. Morphology

- The study of how words are composed of **morphemes** (the smallest meaning-bearing units of a language)
- Two broad classes of morphemes:
  - **Stems**: "main" morpheme of the word, supplying meaning
  - **Affixes**: Bits and pieces that combine with stems to modify their meanings and grammatical functions (prefixes, suffixes, circumfixes, infixes)
    - Unlike
    - Trying
- Multiple affixes
  - Unreadable

Source: Joyce Choi, CSE 842, Michigan State University

---

Ways to Form Words

- **Inflection**: new forms of the same word (usually in the same class)
  - Tense, number, mood, voice marking in verbs
  - Number, gender marking in nominals
  - Comparison of adjectives
- **Derivation**: yield different words in different class
  - Deverbal nominals
  - Denominal adjectives and verbs
- **Compounding**: new words out of two or more other words
  - Noun-noun compounding (e.g., doghouse)
- **Cliticization**: combine a word with a clitic (which acts syntactically like a word but in a reduced form, e.g., I've)

Source: Joyce Choi, CSE 842, Michigan State University

---

English Inflectional Morphology

- Word stem combines with grammatical morpheme
  - Usually produces word of same class
  - Usually serves a grammatical role that the stem could not (e.g., agreement)
    - like -> likes or liked
    - bird -> birds
- Nouns have a simple inflectional morphology: markers for plural and markers for possessives
- Verbs are slightly more complex:

Source: Joyce Choi, CSE 842, Michigan State University
Nominal Inflection

- Nominal morphology
  - Plural forms
    - s or es
    - Irregular forms, e.g., Goose/Geese, Mouse/Mice
  - Possessives
    - children's

Verbal Inflection

- Main verbs (walk, like) are relatively regular
  - -s, -ing, -ed
  - And productive: Emailed, instant-messaged, faxed
  - But eat/ate/eaten, catch/caught/cought
- Primary (be, have, do) and modal verbs (can, will, must) are often irregular and not productive
  - Be: am/is/are/were/was/been/being
- Irregular verbs few (~250) but frequently occurring English verbal inflection is much simpler than e.g. Latin

Regulars and Irregular Verbs

<table>
<thead>
<tr>
<th>Morphological Form Classes</th>
<th>Regularly Inflected Verbs</th>
<th>Irregularly Inflected Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td>merge</td>
<td>eat</td>
</tr>
<tr>
<td>-s form</td>
<td>merges</td>
<td>catches</td>
</tr>
<tr>
<td>-ing participle</td>
<td>merging</td>
<td>eating</td>
</tr>
<tr>
<td>Past form / -ed participle</td>
<td>merged</td>
<td>are</td>
</tr>
<tr>
<td></td>
<td></td>
<td>caught</td>
</tr>
</tbody>
</table>

Source: Joyce Choi, CSE 842, Michigan State University
English Derivational Morphology

- **Word stem** combines with grammatical **morpheme**
  - Usually produces word of **different** class
  - More complicated than inflection

- **Example:** nominalization
  - -ize verbs -> -ation nouns
  - generalize, realize -> generalization, realization

- **Example:** verbs, nouns -> **adjectives**
  - embrace, pity -> embraceable, pitiable
  - care, wit -> careless, witless

Source: Joyce Choi, CSE 842, Michigan State University

---

Morphological Parsing

- Takes a surface input and identifying its components and underlying structure
- Morphological parsing: parsing a word into stem (actually the **base/infinitive form**) and affixes and identifying the parts and their relationships
  - **Base form and features:**
    - goose -> goose + N +SG or goose + V
    - geese -> goose + N + PL
    - gooses -> goose + V +3SG

Source: Joyce Choi, CSE 842, Michigan State University
Inflectional Morphology

- Adds:
  - tense, number, person, mood, aspect
- Word class doesn’t change
- Word serves new grammatical role
- Examples
  - come is inflected for person and number:
    The pizza guy comes at noon.
  - las and rojas are inflected for agreement with manzanas in grammatical gender by -a and in number by -s
    las manzanas rojas (the red apples)

Morphological Analysis Tools

- Porter stemmer
  - A simple approach: just hack off the end of the word!
    Does NOT convert a word to its base form
  - Frequently used in Information Retrieval, but results are pretty ugly!

- WordNet’s morphy()
  - A slightly more sophisticated approach
  - Uses an understanding of inflectional morphology
    - Uses a set of Rules of Detachment
    - Use an Exception List for irregulars
    - Handle collocations in a special way
    - Do the transformation, compare the result to the WordNet dictionary
    - If the transformation produces a real word, then keep it, else use the original word.
    - For more details, see http://wordnet.princeton.edu/man/morphy.7WN.html

Source: Marti Hearst, i256, at UC Berkeley
Some morphy() output

- >>> wntools.morphy('dogs')
  - 'dog'
- >>> wntools.morphy('running', pos='verb')
  - 'run'
- >>> wntools.morphy('corpora')
  - 'corpus'

Source: Marti Hearst, i256, at UC Berkeley

Morphological Analysis Tools

- Very sophisticated programs have been developed
- Use a technique called Two-Level Phonology
  - Has been applied to numerous languages
- Best known: PCKimmo
  - After Kimmo Koskenniemi, based in part on work by Lauri Karttunen in 1983
  - Uses:
    - A rules file which specifies the alphabet and the phonological (or spelling) rules
    - A lexicon file which lists lexical items and encodes morphotactic constraints.
    - http://www.sl.org/pckimmo/
- Commercial versions are available
  - inXight's LinguistX version based on technology developed by Kaplan and others from Xerox PARC (or at least used to be)

Source: Marti Hearst, i256, at UC Berkeley