NLP
Introduction to NLP

Morphology and the Lexicon
Mental Lexicon

• What is the meaning of cat?
  – Its pronunciation?
  – Part of speech?
• What is the meaning of wug?
• What is the meaning of cluvious?
• Compare traftful and traftless.
  – Morphology of these words
• Intuition and productivity
• “Runs”
  – Two interpretations
• Allomorphs
  – cats/oxen, played/swung
• Affixes
Derivational Morphology

• Example
  – “er” (multiple interpretations)

• What do these morphemes mean?
  – prefix, stem, suffix, ending
  – ness, able, ing, re, un, er (adj)
  – JJ \( \rightarrow \) V + “-able”

• Recursion:
  – unconcernednesses

• Ambiguity
  – uncloggable vs. unbelievable
Answer to the Quiz

• Uncloggable
  – unable to be clogged
  – able to be unclogged

• Unbelievable
  – unable to be believed
  – ? able to be unbelieved
Morphological Examples

• Reduplication
  – amigo = friend, amimígo = friends (in Pangasinan) [Rubino 2001]
  – savali = he travels, savavali = they travel (in Samoan)

• Templatic morphology (e.g., Semitic languages):
  – lmd (learn), lamad (he studied), limed (he taught), lumad (he was taught)

• Circumfixes
  – spielen – gespielt (in German)

• Pig Latin
  – appyhay

• Verlan
  – “céfran”, “ripou” (from “l’envers”, “Français”, “pourri”)

• Massa–freakin’–chusetts
  – where can you insert “freakin’” in “education”?
Answer to the Quiz

• The “freakin’” infix is inserted
• … to the left of the syllable that bears the main stress
  edu–freakin’–cation
  * educa–freakin’–tion
  * e–freakin’–ducation
• though there can be exceptions
More Examples

• Clitics
  – l’enfant, cat’s cradle

• Portmanteau words
  – motel, brunch, spork

• Synthetic vs. isolating languages
  – Isolating languages (typically with fixed word order): English, Chinese, Bulgarian, Thai
  – Synthetic languages (high morpheme–per–word ratio): Inuktitut, Ainu, Basque, Lakota

• Fusional vs. agglutinative languages
  – Agglutinative: Turkish, Hungarian, Swahili
  – Fusional: Lithuanian, Hebrew, Latin
Inflectional Morphology

• Many forms
  – Tense, number, person, mood, aspect
  – Five verb forms in English
  – 40+ forms in French
  – Six cases in Russian:
    http://www.departments.bucknell.edu/russian/language/case.html
  – Up to 40,000 forms in Turkish
    • E.g., you cause X to cause Y to ... do Z)
Morphological Analysis

• sleeps = sleep + V + 3P + SG
• done = do + V + PP
## Turkish Vowel Harmony

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unrounded</td>
<td>Rounded</td>
</tr>
<tr>
<td>High</td>
<td>i</td>
<td>ü</td>
</tr>
<tr>
<td>Low</td>
<td>e</td>
<td>ö</td>
</tr>
</tbody>
</table>

- **Back vowels**
  - in the room → odāda
  - at the door → kapıda
- **Front vowels**
  - at home → evde
  - at the lake → gölde
  - on the bridge → köprüde
NACLO Problem

- Turkish
  - by Bozhidar Bozhanov
Given are Turkish words and their English translations:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>güreşçi</td>
<td>wrestler</td>
</tr>
<tr>
<td>B</td>
<td>ikbalsız</td>
<td>unsuccessful</td>
</tr>
<tr>
<td>C</td>
<td>gözcü</td>
<td>sentry, eye doctor</td>
</tr>
<tr>
<td>D</td>
<td>isimsiz</td>
<td>nameless</td>
</tr>
<tr>
<td>E</td>
<td>ormançı</td>
<td>forester</td>
</tr>
<tr>
<td>F</td>
<td>sonsuz</td>
<td>endless</td>
</tr>
<tr>
<td>G</td>
<td>içkici</td>
<td>drunkard</td>
</tr>
<tr>
<td>H</td>
<td>takatsız</td>
<td>lacking strength</td>
</tr>
<tr>
<td>I</td>
<td>barutçu</td>
<td>gunpowder maker</td>
</tr>
<tr>
<td>J</td>
<td>sütsüz</td>
<td>without milk</td>
</tr>
<tr>
<td>K</td>
<td>balıkçı</td>
<td>fisherman</td>
</tr>
<tr>
<td>L</td>
<td>parasız</td>
<td>cashless</td>
</tr>
<tr>
<td>M</td>
<td>mumcu</td>
<td>candlemaker</td>
</tr>
</tbody>
</table>
F1. Two of the above words are formed in a slightly different way from the others because their stems are loans from another language. Identify those two words.

Put their letters here: [ ] [ ] (1.5 points each)
(e.g., D L)

F2. Translate into Turkish (write one letter in each box, starting from the left; it is ok to leave blank boxes after your answer). Use lowercase letters only. Remember that i and i are distinct letters. (2 points each).

milkman: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
blind: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]
F3. Given are the following Turkish words (not loans from another language):

<table>
<thead>
<tr>
<th>dil</th>
<th>language</th>
</tr>
</thead>
<tbody>
<tr>
<td>kalıp</td>
<td>form</td>
</tr>
</tbody>
</table>

Translate into Turkish: (write one letter in each box, starting from the left). Use lowercase letters only (2 points each)

linguist:

mute:

mold maker:

shapeless:

Note: ç sounds like ch in church, c like j in job, ş like sh in shoe. e, i, o, and u are pronounced approximately like in red, reed, rod, and rude, respectively. ö and ü are respectively e and i, pronounced with the lips rounded. ü (written like an “i” but without a dot on top) is like u, pronounced with the lips spread (unrounded).

Turkish is a language from the Turkic group of the Altaic language family. It is spoken by 60 million people in Turkey and roughly 10 million other people around the world.
NACLO Solution

- Turkish
  - www.nacloweb.org/resources/problems/2010/FS.pdf
The two suffixes in the problem have the following meaning:

- **consonant (ç or c) + vowel** is “-er” in English and means “maker of something”
- **s + vowel + z** is “-less” in English and means “without”

Whether the first consonant of the first suffix (“-er”) is ç or c depends on the previous sound:
- if the last sound of the stem is a voiced consonant or a vowel, the first consonant of the suffix is c (also voiced)
- if the last sound of the stem is voiceless consonant, the first consonant of the suffix is ç (also voiceless)

The vowel depends on the last vowel in the stem:
- if last vowel in the stem is a front, unrounded vowel (e, i), the suffix vowel is -i
- if last vowel in the stem is a front, rounded vowel (ö, ü), the suffix vowel is -ü
- if last vowel in the stem is a back, unrounded vowel (a, i), the suffix vowel is -i
- if last vowel in the stem is a back, rounded vowel (o, u), the suffix vowel is -u

To summarize – the vowel in the suffix is the narrow vowel of the same type as the vowel preceding it. This is called **vowel harmony**
F1.
ikbalsız, takatsız – the vowels in the suffixes of these two words do not conform to the rules of vocal harmony and we can assume they are not of Turkish origin.

F2.
milkman - sütçü
speechless – sözsüz

F3.
linguist – dilci
mute - dilsiz
molder – kalıpçı
loose - kalıpsız
Agglutinative Languages

How does English become Turkish?

if we will be able to make ... become strong

... strong become to make be able will if we

... sağlam +laş +tir +abil +ecek +se +k

\[ \downarrow \]

... sağlamlaştırabileceksek
<table>
<thead>
<tr>
<th>ワード</th>
<th>イングリッシュ</th>
</tr>
</thead>
<tbody>
<tr>
<td>アメフト</td>
<td>American Foot(ball)</td>
</tr>
<tr>
<td>アイスクリーム</td>
<td>ice cream</td>
</tr>
<tr>
<td>アイドル</td>
<td>idol</td>
</tr>
<tr>
<td>アパート</td>
<td>apartment</td>
</tr>
<tr>
<td>バイク</td>
<td>bike</td>
</tr>
<tr>
<td>バリアフリー</td>
<td>barrier free</td>
</tr>
<tr>
<td>コンピューター</td>
<td>computer</td>
</tr>
<tr>
<td>デスク</td>
<td>desk (at a news agency)</td>
</tr>
<tr>
<td>ラマ</td>
<td>drama (on TV)</td>
</tr>
<tr>
<td>エレベーター</td>
<td>elevator</td>
</tr>
<tr>
<td>エスカレーター</td>
<td>escalator</td>
</tr>
<tr>
<td>フライドポテト</td>
<td>fried potato (French fries)</td>
</tr>
<tr>
<td>グラス</td>
<td>glass (for drinking)</td>
</tr>
<tr>
<td>ハッピーエンド</td>
<td>happy end(ing)</td>
</tr>
<tr>
<td>ホットケーキ</td>
<td>hotcake (pancake)</td>
</tr>
<tr>
<td>コーヒー</td>
<td>coffee</td>
</tr>
<tr>
<td>クラブ</td>
<td>club</td>
</tr>
<tr>
<td>キーボード</td>
<td>keyboard</td>
</tr>
<tr>
<td>キャンペーン</td>
<td>campaign</td>
</tr>
<tr>
<td>キャップ</td>
<td>cap</td>
</tr>
<tr>
<td>パソコン</td>
<td>pers(onal) com(uter)</td>
</tr>
<tr>
<td>パーソナルコンピューター</td>
<td>personal computer</td>
</tr>
<tr>
<td>レジュメ</td>
<td>resume</td>
</tr>
<tr>
<td>レストラン</td>
<td>restaurant</td>
</tr>
<tr>
<td>リモコン</td>
<td>remo(te) con(trol)</td>
</tr>
<tr>
<td>サラダ</td>
<td>salad</td>
</tr>
<tr>
<td>タバコ</td>
<td>tobacco</td>
</tr>
<tr>
<td>テレビゲーム</td>
<td>television game</td>
</tr>
<tr>
<td>ゼミナール</td>
<td>seminar</td>
</tr>
</tbody>
</table>
Introduction to NLP

Other Levels of Linguistic Analysis
Semantics

• Semantics
  – Lexical semantics and compositional semantics

• Lexical Semantics
  – Hypernyms, hyponyms, antonyms, meronyms and holonyms (part-whole relationship, tire is a meronym of car), synonyms, homonyms
  – Senses of words, polysemous words
  – Collocations
    • white hair, white wine
  – Idioms
    • to kick the bucket

• Compositional Semantics
  – How to understand the meaning of a sentence based on the meaning of its components.
Pragmatics

• The study of how knowledge about the world and language conventions interact with literal meaning.
• Speech acts
• Resolution of anaphoric relations
• Modeling of speech acts in dialogue
Other Areas of Linguistics

• Sociolinguistics
  – interactions of social organization and language.

• Historical linguistics
  – change over time.

• Linguistic typology

• Language acquisition
  – L1 and L2

• Psycholinguistics
  – real-time production and perception of language
NLP