ANLP Week 1 / Lecture 3
Morphology

Adam Lopez

(based on slides by Sharon Goldwater)
Today’s Lecture

• What is morphology, how does it differ across languages, and why does it matter for NLP?
• What’s the difference between a stem, lemma, and affix?
• What are the characteristics of derivational and inflectional morphology?
• What morphological phenomena do we encounter across different languages?
Two plots from last time

![English plot](image)

Frequency vs. Rank for English

![Finnish plot](image)

Frequency vs. Rank for Finnish
# How Many Different Words?

10,000 sentences from the Europarl corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>Different words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>16k</td>
</tr>
<tr>
<td>French</td>
<td>22k</td>
</tr>
<tr>
<td>Dutch</td>
<td>24k</td>
</tr>
<tr>
<td>Italian</td>
<td>25k</td>
</tr>
<tr>
<td>Portuguese</td>
<td>26k</td>
</tr>
<tr>
<td>Spanish</td>
<td>26k</td>
</tr>
<tr>
<td>Danish</td>
<td>29k</td>
</tr>
<tr>
<td>Swedish</td>
<td>30k</td>
</tr>
<tr>
<td>German</td>
<td>32k</td>
</tr>
<tr>
<td>Greek</td>
<td>33k</td>
</tr>
<tr>
<td>Finnish</td>
<td>55k</td>
</tr>
</tbody>
</table>

Why the difference? Morphology.
The word *word* is ambiguous.

- **Word *type***: “10k sentences from English Europarl have 16k different words” (unique strings, lexical items)
- **Word *token***: “English Europarl has 54m words” (possibly repeated instances)

A CAT AND A BROWN DOG CHASED A BLACK DOG:
10 tokens, 7 types.
What is morphology?

- Literally: the study of form
  - From Greek morphē (meaning “form”), and
  - -logy (meaning “study”)
- In linguistics: the study of wordforms and word formation.
  - Structured relationships between words:
    
    play, played, replay, player
    played, walked, jumped
  
  - How the smallest meaningful units of language (morphemes) can be arranged to form word types (morphotactics):
    
    de+salin+ate+ion but not ate+salin+ion+de
Why does morphology matter?

- Information retrieval: return pages with related forms.
- Language modelling: make predictions about unseen words.
- Machine translation and language understanding: signals differences in meaning (might be expressed using word order in other languages).
Why does morphology matter?

THIS IS A WUG.

NOW THERE IS ANOTHER ONE.

THERE ARE TWO OF THEM.

THERE ARE TWO _____.
Why does morphology matter?

Example (Russian):

zhenshina devochke dala knigu
woman+NOM girl+DAT gave book+ACC
‘the woman gave the girl a book’

vs.

zhenshine devochka dala knigu
woman+DAT girl+NOM gave book+ACC
‘the girl gave the woman a book’

A noun’s case marking (a kind of morphology) indicates its role in the sentence, where English uses word order and prepositions.
Morphemes: Stems and Affixes

- Two types of morphemes
  - *Stems* convey core meaning: small, cat, walk
  - *Affixes* can modify meaning: +ed, un+

- Four types of affixes
  - suffix
  - prefix
  - infix
  - circumfix
Stems vs. Lemmas

- **Lemma**: the canonical form or dictionary form of a set of words
  - fly, flies, flew and flying all have the lemma fly.
  - walk, walks, walked and walking all have the lemma walk.
  - walker, walkers have the lemma walker.

- **Stem**: definitions can vary, but often: the part of the word that is common to all its variants
  - stem of produce, production is produc.
  - stem of walk, walks, walked, walking, walker, walkers is walk.
  - Do fly, flies, flew, flying have a common stem fl? Or maybe only fly and flying share a stem: fly. Decision may depend on application.
Suffix

- Plural of nouns
  - $cat+s$

- Comparative and superlative of adjectives
  - $small+er$, $small+est$

- Formation of adverbs
  - $great+ly$

- Verb tenses
  - $walk+ed$

- All inflectional morphology in English uses suffixes
Prefix

• In English: these typically change the meaning
  • Adjectives
    un+friendly
dis+interested
  • Verbs
    re+consider
  • Some language use prefixing much more widely
Infixes

• In English: inserting profanity for emphasis
  abso+bloody+lutely
  fan+fucking+tastic

• Why not:
  ab+bloody+solutely
Circumfix

- No example in English

- German past participle of verb:
  \[ \text{ge+sag+t} \ (\text{German}) \]
Not that easy...

- Affixes are not always simply attached
- In writing, some letters may be changed/added/removed
  - walk+ed
  - frame+d
  - emit+ted
  - carr(−y)+ied
- In speaking, some sounds may be changed/added/removed
  - Compare the final sound: cats [s] vs dogs [z] vs foxes [əz]
Morphological processes other than affixation

• Root and pattern

<table>
<thead>
<tr>
<th>root</th>
<th>pattern</th>
<th>phonological form</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ktb</td>
<td>CaCaC</td>
<td>katav</td>
<td>'wrote' (Hebrew)</td>
</tr>
<tr>
<td>ktb</td>
<td>hiCCiC</td>
<td>hixtiv</td>
<td>'dictated'</td>
</tr>
<tr>
<td>ktb</td>
<td>miCCaC</td>
<td>mixtav</td>
<td>'a letter'</td>
</tr>
</tbody>
</table>

• Reduplication

anak ‘child’  anak-anak ‘children’  (Indonesian)
Irregular Forms

- Some words have irregular forms:
  - is, was, been
  - eat, ate, eaten
  - go, went, gone

- Irregular forms tend to be the most frequent (and vice versa)
- Probably a cognitive explanation: regular morphology allows us to express new meanings without memorizing new words. (But then: why irregulars?)
Inflection and derivation
Inflectional vs. Derivational Morphology

- So far, distinctions are mainly about **form**: where does the morpheme go, what does it look like?
- We can also distinguish more by **function**: inflection or derivation?
- Inflectional morphology typically
  - does not change basic meaning or part of speech
  - expresses grammatical features or relations between words
  - applies to all words of the same part of speech
Inflectional Morphology

• In English, we inflect
  • *nouns* for count (plural: +s) and for possessive case (+’s)
  • *verbs* for tense (+ed, +ing)
    ▶ and 3rd person singular present form (+s) to agree with subject
      I swim, you swim, she swims
  • *adjectives* in comparative (+er) and superlative (+est) forms.

• In German, we inflect
  • *nouns* for count and case
  • *verbs* for tense, person, and count
  • *adjectives* for count, case, gender, and definiteness
  • *determiners* for count, case and gender
Forms of the German the

<table>
<thead>
<tr>
<th>Case</th>
<th>Singular</th>
<th></th>
<th>Plural</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
<td>fem.</td>
<td>n.</td>
<td>male</td>
</tr>
<tr>
<td>nominative (subject)</td>
<td>der</td>
<td>die</td>
<td>das</td>
<td>die</td>
</tr>
<tr>
<td>genitive (possessive)</td>
<td>des</td>
<td>der</td>
<td>des</td>
<td>der</td>
</tr>
<tr>
<td>dative (indirect object)</td>
<td>dem</td>
<td>der</td>
<td>dem</td>
<td>den</td>
</tr>
<tr>
<td>accusative (direct object)</td>
<td>den</td>
<td>die</td>
<td>das</td>
<td>die</td>
</tr>
</tbody>
</table>

Phrase/role: [The A]/s put [the B]/o [of the C]/p [on the D]/io

Not only many different forms, but each form is highly ambiguous.
Inflectional vs. Derivational Morphology

• **Inflectional morphology typically**
  • does not change basic meaning or part of speech
  • expresses grammatical features or relations between words
  • applies to all words of the same part of speech

• **Derivational morphology**
  • may change the part of speech or meaning of a word
  • is not driven by syntactic relations outside the word
  • may be “picky”: intractibil+ity but not intractible+ness
  • applies closer to the stem; whereas inflection occurs at word edges: govern+ment+s, centr+al+ize+d
Derivational Morphology

• Changing the part of speech, e.g. noun to verb
  \[\text{word} \rightarrow \text{wordify}\]

• Is it a real word?

• Consulting Google (a few years ago):
  • 8,840 hits: e.g., \text{wordify mugs, tshirts and magnets}

• Google now returns over 225k hits. (Why?)
Derivational Morphology

- Changing the verb back to a noun
  \[\text{wordify} \rightarrow \text{wordification}\] (8k hits on Google)

- A person/thing who engages in \textit{wordification}
  \[\text{wordification} \rightarrow \text{wordificator}\] (was 8 hits, now 9k: another app!)

- A person/thing who \textit{wordifies}
  \[\text{wordify} \rightarrow \text{wordifier}\] (1700 hits on Google)

- What is the difference between a \textit{wordifier} and a \textit{wordificator}?

- Turning \textit{wordification} into an ideology:
  \[\text{wordification} \rightarrow \text{wordificationism}\]
Derivational Morphology

• An adherent of wordificationism
  \( \text{wordificationism} \rightarrow \text{wordificationist} \)

• Used to have 0 hits on Google, now you get these slides!
• We created a new word!
Compounds

• Creating new words by merging multiple words
• (Somewhat) rare in English
  
  home work → homework
  
  web site → website

• More common in other languages (like German)
Morphemes can be ambiguous

- **Lexical ambiguity** due to ambiguous root morpheme:
  - bank (financial institution) vs. bank (land alongside a river)
- Affixes can also be ambiguous:
  - +s indicates plural on nouns (dogs)
  - +s indicates present tense on 3rd person singular verbs (swims)
- Consider she+s. Does it mean she is or she has?
- Affix ambiguity is not always resolved by the root morpheme:
  - (un+tie)+able: Able to be untied
  - un+(tie+able): Unable to be tied
Morphological variation across languages
Morphology differs across languages

• Languages with few morphemes per word: **isolating**
• Languages with many morphemes per word: **synthetic**
• Often a trade-off between morphology and word order (syntax)
  • Some languages have no verb tenses
    → use explicit time references (**yesterday**)
  • Case inflection determines roles of noun phrase
    → use fixed word order instead
    → use prepositional phrases instead of cased noun phrases
• Examples from the World Atlas of Language Structures (**wals.info**)
  • prefixes vs. suffixes
  • cases (zero to more than ten)
  • past tense remoteness distinctions