Foundations of Natural Language Processing
Lecture 11a
Introduction to Morphology

Alex Lascarides
(slides based on those of Shay Cohen)

School of informatics

Last Time

- A range of ML methods that are useful for NLP
- Today: **Morphology**:
  Words aren't the smallest unit!

Morphology

This lecture:

- The problem (some examples)

Later. . .

- Finite State Transducers (FSTs)
- Using FSTs to tackle morphology parsing and generation

Morphology

- Morphology is the study of the *structure of words*
  - English has relatively impoverished morphology.
  - Languages with rich morphology: Turkish, Arabic, Hungarian, Korean . . .

- For example, Turkish is an *agglutinative* language: tends to express concepts in complex words constructed by concatenating *morphemes*, each expressing some simple (component) concept:

evlerinizden: "from your houses" morphemes:	ev-	ler-	iniz-	den
	house	plural	your	from

- This lecture will mostly discuss how to build an English morphological analyser.

Morphology

'Whole' words constructed by combining:

1. Stems (*house, combine, eat, walk,. . .*)
   The 'dictionary' bit

2. Affixes (prefixes, suffixes, infixes and circumfixes)
   grammar parts.
   The type of affix is determined by where it goes with respect to the stem.

Prefix	before the stem
Suffix	after the stem
Infix	middle of the stem
Circumfix	'reduction' of the stem

Four methods to combine them

Inflection (stem + grammar affix): no change to grammatical category
(*walk* $\rightarrow$ *walking*)

Derivation (stem + grammar affix): change to grammatical category
(*combine* $\rightarrow$ *combination*)

Compounding (stems together): *doghouse*

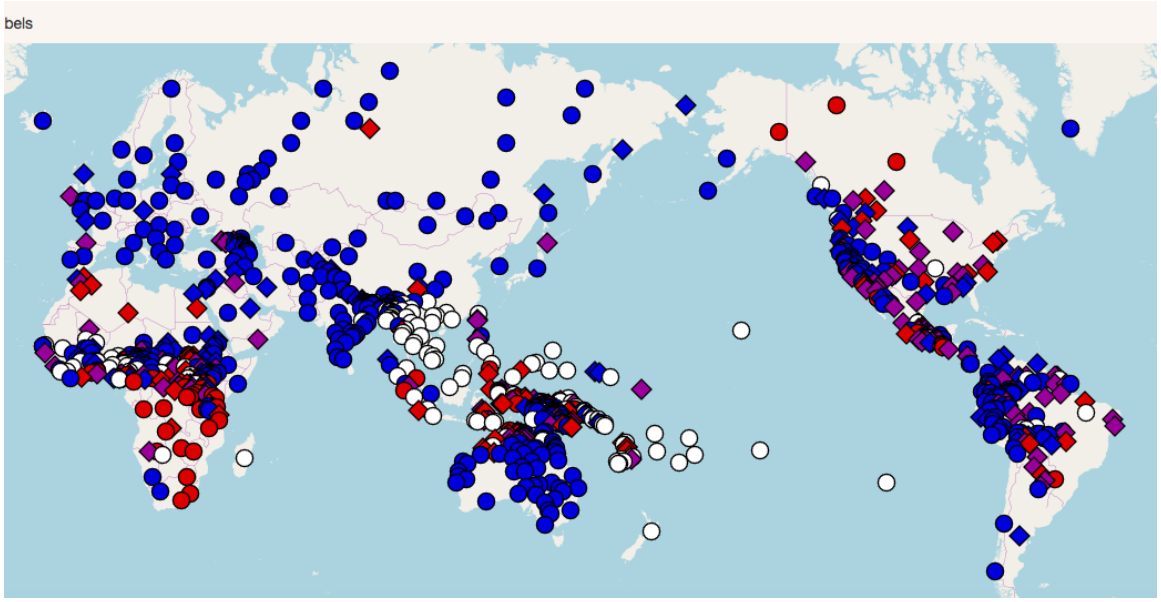
Cliticization: *I've, we're, he's*

Morphology can be concatenative
or non-concatenative (e.g. templatic morphology as in Arabic)

Suffixing versus Prefixing

Values

○	Little affixation	141
●	Strongly suffixing	406
◆	Weakly suffixing	123
◆	Equal prefixing and suffixing	147
◆	Weakly prefixing	94
●	Strong prefixing	58



Different inflection in different languages

- English: nouns are inflected for number; verbs for person and tense
  - *book* (1)/ *books* (> 1)
  - *You read* (2nd pers. present or past)
    she reads (3rd pers. present)
    she read (3rd pers. past)

- German: nouns inflected for number and case:

	Singular	Plural
Nominative	das Haus	die Häuser
Genitive	des Hauses	der Häuser
Dative	dem Haus / dem Hause	den Häusern
Accusative	das Haus	die Häuser

- Spanish: inflection depends on gender (*el sol*/*la luna*)

- Luganda: nouns have ten genders!

Examples: Agglutination and compounding

- *ostoskeskuksessa*
  ostos#keskus+N+Sg+Loc:in
  shopping#center+N+Sg+Loc:in
  'in the shopping center' (Finnish)

- *qangatasuukkuvimmuuriaqalaaqtunga*
  "I'll have to go to the airport" (Inuktitut)

- *Avrupallatramadklarmzdanmsnzcasna*
  "as if you are reportedly of those of ours that we were unable to Europeanize" (Turkish)

In the most extreme examples, the meaning of the word is the meaning of a sentence!

Morphological parsing: the problem

- English has concatenative morphology. Words can be made up of a main stem plus one or more affixes carrying grammatical information.

Surface form:	cats	walking	smoothest
Lexical form:	cat+N+PL	walk+V+PresPart	smooth+Adj+Sup

- Morphological parsing is the problem of extracting the lexical form from the surface form. (For ASR, this includes identifying the word boundaries.)

- We should take account of:
  - Irregular forms (e.g. goose $\rightarrow$ geese)
  - Systematic rules (e.g. 'e' inserted before suffix 's' after s,x,z,ch,sh: fox $\rightarrow$ foxes, watch $\rightarrow$ watches)
  - Things that look like affixes but aren't (*proactive* vs. *protect*)
  - Blocking: **semi**-productivity of morphological rules: N+*ful* $\mapsto$ Adj (*graceful*, *pityful* . . .), but **intelligenceful* (intelligent), **gloryful* (glorious). . .

Summary

- Words have structure:
  - Stems, prefixes, suffixes, stems together, cliticization
- Morphology is the study of that structure.
- Morphology can affect a word's grammatical category and meaning
- Important to know its lexical form (e.g., cat+N+PL) as well as its surface form (cats).
- Morphology varies radically across languages
- Challenging problem because of:
  - Irregularity
  - Some stems have parts that look like affixes, but aren't!
  - Blocking and semi-productivity.