Why isn’t the plural of Smurf Smurves?

One house, two hice.

For some reason, nobody wants to talk to us!

CS 671 NLP
MORPHOLOGY

amitabha mukerjee
iit kanpur

Source: urbanblah
Levels of Linguistic Analysis

**Phonology**

↔ /mohallekaeklaRkA/

**Morphology**

↔ गोल्गोले का एक लड़का

**Syntax**

mohalle ka ek laRkA

**Semantics**

Boolean Logic:

∃x ∃y boy(x) ^ loc(y)^ lives-at(x,y)]

Alternate: Imagistic

loc:y
Syntax vs Morphology

• **Syntax**: how words can be assembled into phrases / sentences:
  - *I found an unopened bottle of wine*
  - *I found a bottle unopened of wine*

• **Morphology**: internal form of words
  - *unopened* – not *openuned* or any other order

• But this distinction is not crisp (since notion of “morpheme” or “word” is graded) → **Morphosyntax**
Syntax / Semantics divide

- Traditional view:
  - Syntax / Morphology: Deals with the form of words (the phonology). Different from
  - Semantics: The study of the meaning for these forms

- Cognitive view:
  Semantics is involved in all composition operations.
Morphemes?

Traditional view:

- **Morphemes**: meaning-carrying units, but not independent

- Morphemic decomposition can be problematic – e.g.,
  
  *take* $\rightarrow$ *took*;

  Hindi: भीखम राम ने उनको छुड़वाया

  *chhuR* $\rightarrow$ *chhuRwaya*

  release causative; caused to release
Morpheme examples

- अपहरणकर्ता = नि- [रीक्षा] -क

- bound / free morphemes:
  -क vs -कर्ता (e.g. अपहरणकर्ता)

- Morphemes often cause changes to the stem
  - bAngla: kin- , buy
    - Ami kinAm
    - uni kenen
    - kenAkATA
    - I buy+PAST
    - he (honorific) buy+PRES
    - buying (noun)
Morpheme positions

- **prefix**
  - dis- (dislike), mis- (misunderstood)
  - com-, de-, dis-, in-, re-, post-, trans-, ...

- **suffix**
  - -able (movable) / -ly (quickly)
  - -tion, -ness, -ate, -ful, ...

- **infix**
  - arundhati “leftist” roy
  - छुड़ाया chuRAyA → छुड़वाया chhurwAyA

- **circumfix**
  - Rare in English – e.g. “a-jumping we shall go”
  - Hindi? (mostly changes stem as well)
Agglutinative: Finnish Noun Declension

**talo 'house'**
- talo 'the-house'
- talo-ni 'my house'
- talo-ssa 'in the-house'
- talo-ssa-ni 'in my house'
- talo-i-ssa 'in the-houses'
- talo-i-ssa-ni 'in my houses'

**kaup-pa 'shop'**
- kaup-pa 'the-shop'
- kaup-pa-ni 'my shop'
- kaup-a-ssa 'in the-shop'
- kaup-a-ssa-ni 'in my shop'
- kaup-o-i-ssa 'in the-shops'
- kaup-o-i-ssa-ni 'in my shops'
Stemming (baby lemmatization)

- Assumption: surface form = root . affix

- Reduce a word to the main morpheme

  automate
  automates
  automatic
  automation

  run
  runs
  running

- Widely used in Information Retrieval
Porter Stemmer (1980)

- Most common algorithm for stemming English
  - Results suggest it’s at least as good as other stemming options
- Multiple sequential phases of reductions using rules, e.g.
  - sses → ss
  - ies → i
  - ational → ate
  - tional → tion

- [http://tartarus.org/~martin/PorterStemmer/](http://tartarus.org/~martin/PorterStemmer/)
Stemming example

Candidate = candid + ate

This is a poorly constructed example using the Porter stemmer.

This is a poorli construct example us the Porter stemmer.

http://maya.cs.depaul.edu/~classes/ds575/porter.html
Code:
http://snowball.tartarus.org/algorithms/english/stemmer.html
Inflections and Derivations

- **Inflection**: e.g. *sing* → *sang* ; *cat* → *cats*
  - variation in form due to tense, person, etc.
  - does not change primary meaning,
  - same part-of-speech
  - applies to nearly entire class of lexemes

- **Derivation**: e.g. *sing* → *singer*
  - changes meaning, changes part-of-speech
  - Like much in grammar, not very crisp distinction
  - e.g *cyclic* → *cyclical* = derivation
  - treat as new word
Productive Morphemes

- A morpheme is productive if it applies to all words of a given type.
- Inflections – almost fully productive
- Derivations – very limited
Inflections

- **paradigm**: set of inflections in given grammar
  - person (1 2 3)
  - number (singular *sg*, plural *pl*), and
  - tense (present, simple past):

<table>
<thead>
<tr>
<th></th>
<th>sg</th>
<th>1-sg</th>
<th>2-sg</th>
<th>3-sg</th>
</tr>
</thead>
<tbody>
<tr>
<td>pres</td>
<td>i sing,</td>
<td>you sing,</td>
<td>[s]he sings,</td>
<td></td>
</tr>
<tr>
<td>past</td>
<td>i sang,</td>
<td>you sang,</td>
<td>[s]he sang,</td>
<td></td>
</tr>
</tbody>
</table>

**paradigm:**

*sing, v.*

<table>
<thead>
<tr>
<th></th>
<th>pl</th>
<th>1-pl</th>
<th>2-pl</th>
<th>3-pl</th>
</tr>
</thead>
<tbody>
<tr>
<td>pres</td>
<td>we sing,</td>
<td>you sing,</td>
<td>they sing</td>
<td></td>
</tr>
<tr>
<td>past</td>
<td>we sang,</td>
<td>you sang,</td>
<td>they sang</td>
<td></td>
</tr>
</tbody>
</table>
Sanskrit Morphology

• Sanskrit paradigms - *pratyaya* :– six types
  • *sup*- nominal inflections, (*subanta*)
  • *tin*- verb inflections, temporal and modal (*tinanta*)
  • *krt* – noun formation e.g. $kr^\wedge + \text{tavya} = \text{kartavya}$
    [do + to-be-done = duty]
  • *taddhita*- nouns from nouns : secondary forms
  • *dhAtu*- verbal endings
  • *stri*- gender formations

• Both inflections and derivations
Noun paradigm: karakas (sup-)

Masculine, singular, -a forms

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>devas</td>
<td>nominative</td>
</tr>
<tr>
<td>2</td>
<td>devam</td>
<td>accusative</td>
</tr>
<tr>
<td>3</td>
<td>devena</td>
<td>Instrumental</td>
</tr>
<tr>
<td>4</td>
<td>devāya</td>
<td>dative</td>
</tr>
<tr>
<td>5</td>
<td>devāt</td>
<td>ablative</td>
</tr>
<tr>
<td>6</td>
<td>devasya</td>
<td>genitive</td>
</tr>
<tr>
<td>7</td>
<td>deve</td>
<td>locative</td>
</tr>
<tr>
<td>8</td>
<td>deva</td>
<td>vocative</td>
</tr>
</tbody>
</table>

B. Blake, Case 2001
Inflections

- Languages vary in richness of paradigm
  - English: *to love* four shapes: *love, loves, loved, loving*
  - Latin: *amo* : over a hundred shapes [Sanskrit: ~ 90]
  - Chinese : almost invariant [Analytic]
  - Arabic : *shakara* 'to thank' - can generate 2552 forms
  - Indo-Aryan: despective / honorific forms *tu jA / Ap jAiye*;

- Paradigms for noun / adjective etc.
  - Inflections can apply to other word categories
  - E.g. case: *rAm ne khAnA khAyA* :
    - morpheme *ne* marks the noun *rAm* as having a subject relation to the head of the phrase, *khA*
Derivations

(Lexical Morphology)

e.g. endanger from en- + danger
Derivations: Word formation

- Inflection vs Derivations: very ancient distinction
  - originated by sakaTAyana (8th c. bce) : most nouns derived from some verb root (dhAtu)
    e.g. join → joint
  - Yaska’s nirukta [etymology] (6th c. bce),
  - pAniNi’s aShTAdhyAyI (5th c.) – argues against this view. Distinguishes Inflections (pratyaya) from derivations (krit)

- Derivations: krit : noun-forms from the verb
  - kr^ + -tavya → kartavya [do + to-be-done = duty]
    (similar to do+ -able → doable)
Derivations

- e.g. **ungentlemanly**: un + gentle + man + ly

- not all lexemes of a class will take all these particles, nor will they have the same meaning.

- how to break up (parse) the lexeme?
  - [ [un+gentle] + man ] + ly
  - [un + [gentle + man] ] + ly

  many interpretations are possible
Derivations: Parsing

- Differing parses $\rightarrow$ different semantics:
- e.g. unlockable
  - “can’t be locked” or “can be unlocked”?

Huddleston & Pullum 05
Derivations: Ambiguity

- Semantics: not fully systematic – e.g. anomalous usage of *un*-:
  - *loosen* same as *unloosen*
Semantics of morphemes

- **inflections:**
  - e.g. “-ed” : past tense = events in the past
    - *The course started last week.*
  - *But:* often does not refer to past, e.g.:
    - *I thought the course started next week.*
    - *If the course started, everyone would be pleased.*

- past time = **primary** or most common characteristic
- many other interpretations possible (in many languages)
  - → *past tense* = grammatical form, varied semantics
Semantics of composition

- **derivations:**
  - e.g. “-er” : usually agentive – *builder, writer, teacher*
  - But may be instrumental – e.g. *cooker*
    - However, meaning is constrained (not arbitrary)

- **compounds:** composed from multiple lexemes
  - *doghouse, darkroom* (endocentric, tatpurusha) : ‘house’, ‘room’ is the head
  - *redcoat, Hindi: nllakanTha* (exocentric, bahuvrihi) : refers to neither red nor coat
Computational Morphology
Computational Analysis

• [Harris 1955]

/hiyzkwikor/ He's quicker
will have the segmentation: /hiy.z.kwik.or/;
→ To be done "purely by comparing this phonemic sequence with the phonemic sequences of other utterances."

• [Keshava Pitler 06]: Based on transition frequencies –
  How many starting syllables are un-?
  • Best results for English - 2006 PASCAL challenge
Computational Analysis

- [Goldsmith 01]
  Information-Theoretic ideas - Minimum Description Length

Which “signature” (pattern) will result in the most compact description of the corpus?

<table>
<thead>
<tr>
<th>Signature</th>
<th>Example</th>
<th>Stem # (type)</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL.ed.ing</td>
<td>betray betrayed betraying</td>
<td>69</td>
<td>864</td>
</tr>
<tr>
<td>NULL.ed.ing.s</td>
<td>remain remained</td>
<td>14</td>
<td>516</td>
</tr>
<tr>
<td></td>
<td>remaining remains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NULL.s.</td>
<td>cow cows</td>
<td>253</td>
<td>3414</td>
</tr>
<tr>
<td>e.ed.es.ing</td>
<td>notice noticed notices</td>
<td>4</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>noticing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computational Analysis

- [Dasgupta & V.Ng 07]
  - Simple concatenation not enough for more agglutinated languages.
  - Attempt to discover root word form. \((\text{denial} \rightarrow \text{deny})\)
  - Assumption: if compound word is common, then root word will also: Word-Root Frequency Ratios (WRFR)

<table>
<thead>
<tr>
<th>Correct Parses</th>
<th>Incorrect Parses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Root</strong></td>
</tr>
<tr>
<td>bear-able</td>
<td>bear</td>
</tr>
<tr>
<td>attend-ance</td>
<td>attend</td>
</tr>
<tr>
<td>arrest-ing</td>
<td>arrest</td>
</tr>
<tr>
<td>sub-group</td>
<td>group</td>
</tr>
<tr>
<td>re-cycle</td>
<td>cycle</td>
</tr>
<tr>
<td>un-settle</td>
<td>settle</td>
</tr>
</tbody>
</table>
# Computational Analysis

- [Dasgupta & V.Ng 07]

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Bengali</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>P</td>
<td>R</td>
<td>F</td>
<td></td>
<td>A</td>
<td>P</td>
<td>R</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Linguistica</td>
<td>68.9</td>
<td>84.8</td>
<td>75.7</td>
<td>80.0</td>
<td></td>
<td>36.3</td>
<td>58.2</td>
<td>63.3</td>
<td>60.6</td>
<td></td>
</tr>
<tr>
<td>Morphessor</td>
<td>64.9</td>
<td>69.6</td>
<td>85.3</td>
<td>76.6</td>
<td></td>
<td>56.5</td>
<td>89.7</td>
<td>67.4</td>
<td>76.9</td>
<td></td>
</tr>
<tr>
<td>Basic induction</td>
<td>68.1</td>
<td>79.4</td>
<td>82.8</td>
<td>81.1</td>
<td></td>
<td>57.7</td>
<td>79.6</td>
<td>81.2</td>
<td>80.4</td>
<td></td>
</tr>
<tr>
<td>Relative frequency</td>
<td>74.0</td>
<td>86.4</td>
<td>82.5</td>
<td>84.4</td>
<td></td>
<td>63.2</td>
<td>85.6</td>
<td>79.9</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td>Suffix level similarity</td>
<td>74.9</td>
<td>88.6</td>
<td>82.3</td>
<td>85.3</td>
<td></td>
<td>66.1</td>
<td>89.7</td>
<td>78.8</td>
<td>83.9</td>
<td></td>
</tr>
<tr>
<td>Allomorph detection</td>
<td><strong>78.3</strong></td>
<td>88.3</td>
<td>86.4</td>
<td><strong>87.4</strong></td>
<td></td>
<td><strong>68.3</strong></td>
<td>89.3</td>
<td>81.3</td>
<td><strong>85.1</strong></td>
<td></td>
</tr>
</tbody>
</table>