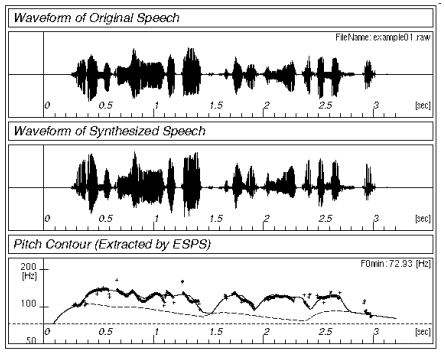
Source: urbanblah

CS 671 NLP MORPHOLOGY

amitabha mukerjee
iit kanpur

Levels of Linguistic Analysis

Phonology



⇔ /mohallekaeklaRkA/

Morphology /mohallekaeklaRkA/ ⇔ मोहल्ले का एक लड़का

Syntax mohalle ka ek laRkA

मोहल्ले का एक **लड़का**

loc np

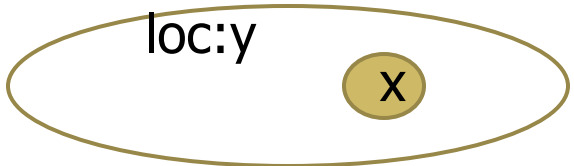
मोहल्ले **का** एक **लड़का**

Semantics

Boolean Logic:

$\exists x \exists y$ boy(x) ^ loc(y)^ lives-at(x,y)]

Alternate: Imagistic



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 → took;

Hindi: *भीखम राम ने उनको छुड़वाया*

chhuR → chhuRwaya

release causative; caused to release

Morpheme examples

- अपहरणकर्ता = नि- [रीक्ष] -क
- prefix suffix

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

- Morphemes often cause changes to the stem
  - bAngla: kin- , buy

Ami kinlAm	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

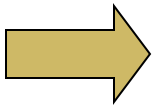
talo 'house'	*kaup-pa* 'shop'
talo 'the-house’	kaup-pa 'the-shop'
talo-ni 'my house'	kaup-pa-ni 'my shop'
talo-ssa 'in the-house'	kaup-a-ssa 'in the-shop'
talo-ssa-ni 'in my house’	kaup-a-ssa-ni 'in my shop'
talo-i-ssa 'in the-houses’	kaup-o-i-ssa 'in the-shops'
talo-i-ssa-ni 'in my houses’	kaup-o-i-ssa-ni 'in my shops'

Stemming (baby lemmatization)

- Assumption : surface form = root . affix

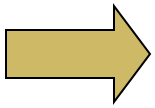
- Reduce a word to the main morpheme

automate
automates
automatic
automation



automat

run
runs
running



run

- 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 $\rightarrow$ ss
  - ies $\rightarrow$ i
  - ational $\rightarrow$ ate
  - tional $\rightarrow$ tion
- 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):

paradigm:
sing, v.

sg	1-sg	2-sg	3-sg
pres	i sing,	you sing,	[s]he sings,
past	i sang,	you sang,	[s]he sang,

pl	1-pl	2-pl	3-pl
pres	we sing,	you sing,	they sing
past	we sang,	you sang,	they sang

Sanskrit Morphology

- Sanskrit paradigms - *pratyaya* :– six types
  - *sup-* nominal inflections, (*subanta*)
  - *tin-* verb inflections, temporal and modal (*tin*anta)
  - *krt –* noun formation e.g. *kr^ + tavya = 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

1	*devas*	nominative	kartr^
2	*devam*	accusative	karman
3	*devena*	Instrumental	karaNa
4	*devāya*	dative	sampradāna
5	*devāt*	ablative	apādāna
6	*devasya*	genitive	samvandha
7	*deve*	locative	adhikaraNa
8	*deva*	vocative	

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: despictive / 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

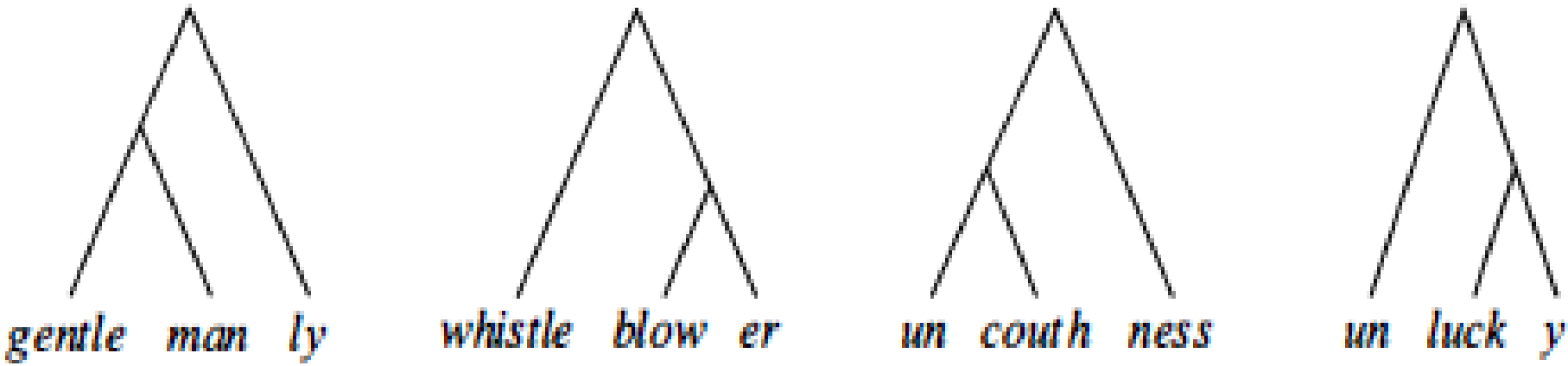
- e.g. **ungentlemanly:** un + gentle + man + ly

- 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

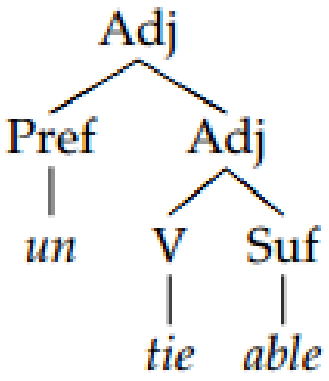


- Differing parses → different semantics :
- e.g. unlockable

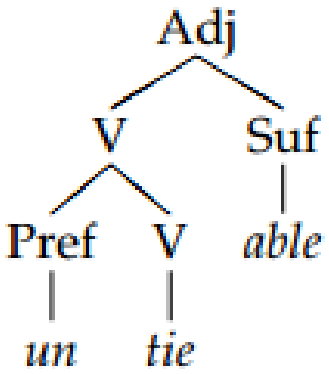
  “can’t be locked” or “can be unlocked”?

Huddleston & Pullum 05

Derivations : Ambiguity



This knot is loose –
it's easily **untieable**



This rope is too slippery –
it's **untieable**

- 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: *nIlakanTha* (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 results in the most compact description of the corpus?

Signature	Example	Stem # (type)	Counts Token
NULL.ed.ing	betray betrayed betraying	69	864
NULL.ed.ing.s	remain remained remaining remains	14	516
NULL.s.	cow cows	253	3414
e.ed.es.ing	notice noticed notices noticing	4 62	

Computational Analysis

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

Correct Parses			Incorrect Parses		
Word	**Root**	**WRFR**	**Word**	**Root**	**WRFR**
bear-able	bear	0.01	candid-ate	candid	53.6
attend-ance	attend	0.24	medic-al	medic	483.9
arrest-ing	arrest	0.06	prim-ary	prim	327.4
sub-group	group	0.0002	ac-cord	cord	24.0
re-cycle	cycle	0.028	ad-diction	diction	52.7
un-settle	settle	0.018	de-crease	crease	20.7

Computational Analysis

- [Dasgupta & V.Ng 07]

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