CS626: Speech, Natural Language Processing and the Web

Part of Speech Tagging (discriminative models and role of morphology)

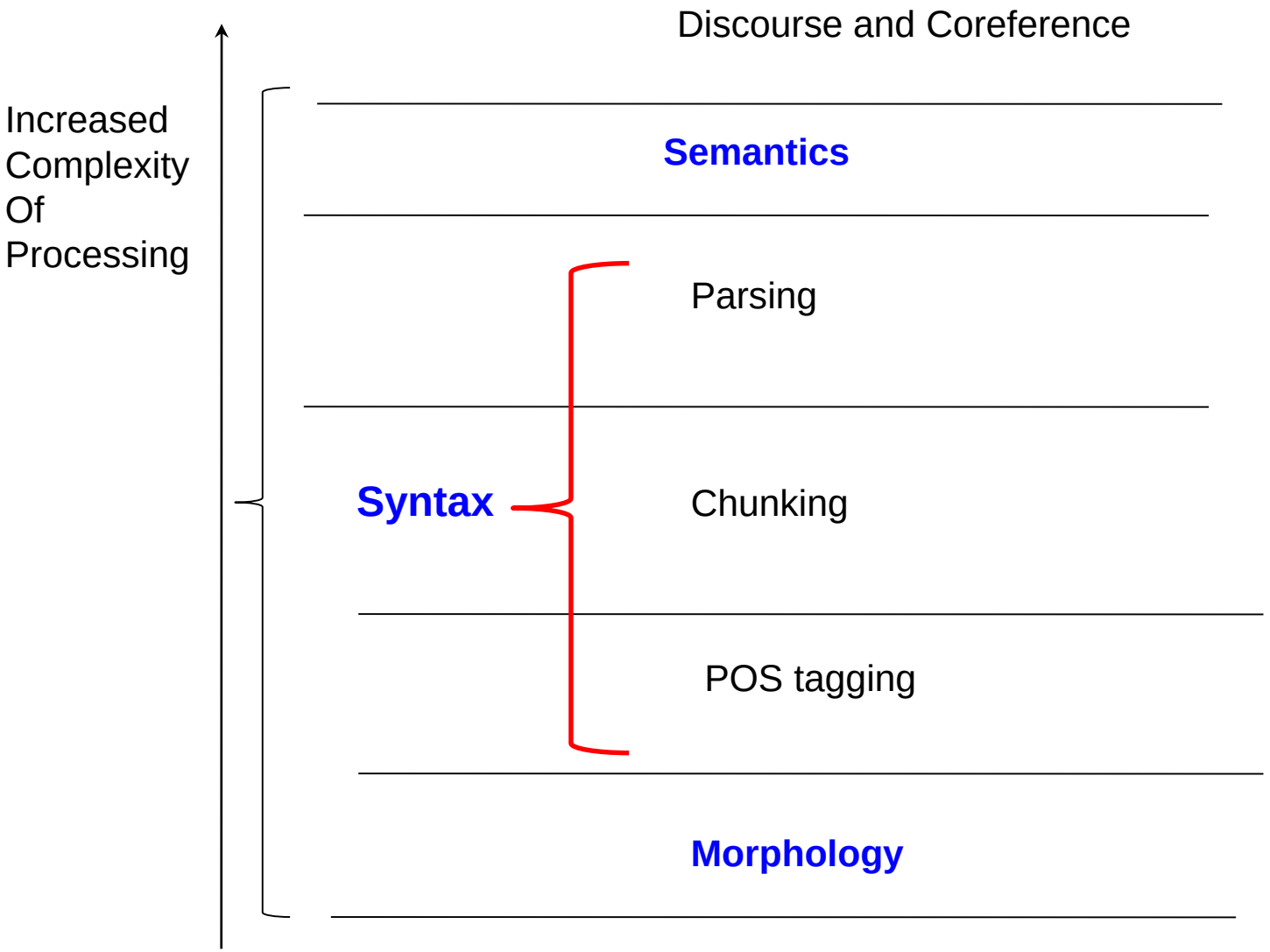
Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week 4 of 15th August, 2022

NLP Layers


Discourse and Coreference
Increased
Complexity
Of
Processing
Semantics
Parsing
Syntax
Chunking
POS tagging
Morphology

What does POS tagging Facilitate

Facilitates Chunking: small phrases called **Chunks**

- given the sentence

  The brown fox sat in front of the fence

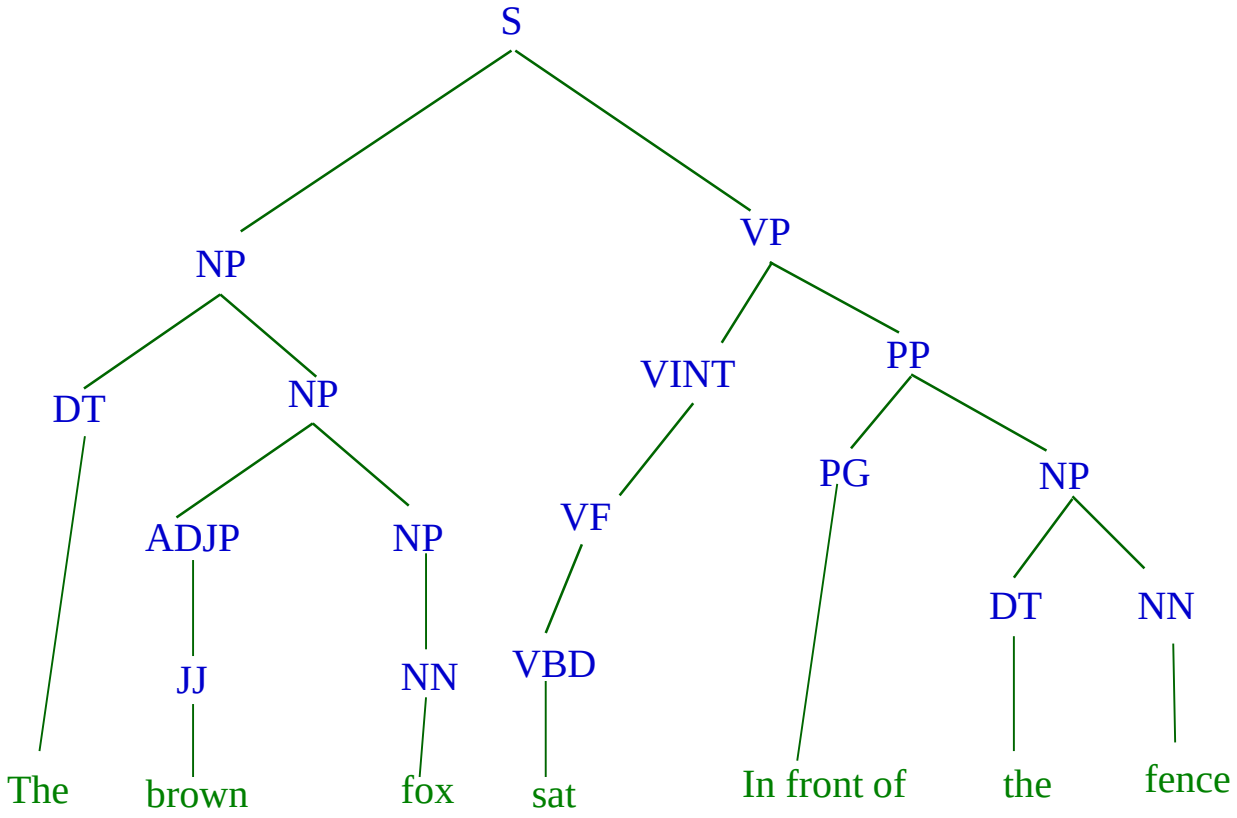
- POS tagged sequence as

  The_DT brown_JJ fox_NN sat_VBD
  in_IN front_NN of_IN the_DT fence_NN

Chunked sequence as

The_DT_B_{NC} brown_JJ_I_{NC} fox_NN_I_{NC}
sat_VBD_B_{VC} in_IN_B_{PC} front_NN_I_{PC}
of_IN_I_{PC} the_DT_B_{NC} fence_NN_I_{NC}

Deep Parse Tree of *the brown fox sat in front of the fence*



Grammar rules

- S → NP VP
- NP → DT NP | ADJP NP | PP NP | NNS | NN
- ADJP → ADJP JJ | JJ
- PP → PG NP | P NP
- PG → ‘in front of’ | ‘in lieu of’ | ‘with respect to’ | …
- P → ‘in’ | ‘with’ | ‘by’ | …
- NN → ‘fox’ | ‘fence’ | …
- JJ → ‘brown’ | …
- DT → ‘a’ | ‘an’ | ‘the’ | …

- VP → VT NP | VINT PP
- VT → VXG VF | VF
- VINT → VXG VF | VF
- VXG → VXG VX | VX
- VF → VB | VBD | …
- VX → ‘am’ | ‘is’ | ‘shall’ | …
- VB → ‘go’ | ‘see’ | …
- VBD → ‘sat’ | ‘went’ | …
- NN → ‘fox’ | ‘fence’ | …

Discriminative Labelling

Motivation

- HMM based POS tagging cannot handle “free word order” and “agglutination” well
- If *adjective after noun* is equally likely as *adjective before noun*, the transition probability is no better than uniform probability which has high entropy and is uninformative.
- When the words are long strings of many morphemes, POS tagging w/o morph features is highly inaccuarte.

Modelling

$$\prod_{i=0}^{n+1} [P(t_i \mid F_i)]$$

Feature Engineering

- *A. Word-based features*

  f_{21} – dictionary index of the current word ('foxes'): integer

  f_{22} – -do- of the previous word ('brown'): integer

  f_{23} – -do- of the next word ('jumped'): integer

- *B. Part of Speech (POS) tag-based feature*

  f_{24} – index of POS of previous word (here JJ): integer

Feature engineering cntd.

- *C. Morphology-based features*

  - f_{25}– does the current word (‘foxes’) have a noun suffix, like ‘s’, ‘es’, ‘ies’, etc.: 1/0- here the value is
  - f_{26}– does the current word (‘foxes’) have a verbal suffix, like ‘d’, ‘ed’, ‘t’, etc.: 1/0- 0
  - f_{27} and f_{28} for ‘brown’ like for ‘foxes
  - f_{29} and $f_{2,10}$ for ‘jumped’ like for ‘foxes; here $f_{2,10}$ is 1 (jumped has ‘ed’ as suffix)

A note of morph features (1/2)

- Morphology features can be fairly open ended, large in number and complex depending on the language under consideration.
- Dravidian languages, Tibeto-Burman languages, Arabic, Hungarian, Turkish, Finnish and so on are morphologically complex.

A note of morph features (1/2)

- Used with dexterity, they can disambiguate POS tags with very high degree of certainty.

- For example, the '*unnu*' suffix in the Malayalam word '*ceyy-unnu*': English-'*does, is doing*' is a sure-shot identifier of verb POS (VBS).

A note on morphology

Typology of languages wrt morphology (1/2)

- Languages of the world fall at various points in the *analytic-synthetic* spectrum.
- **Analytic** languages: morphemes largely separate from one another
- **Synthetic** languages: join the morphemes.
- **Morphemes**: smallest meaning-bearing units forming a word.
  - '*quickly*': '*quick*' and '*ly*'.

Typology of languages wrt morphology

- No language is completely analytic or completely synthetic.
- For example, to express future tense of ‘go’ activity, English uses two morphemes separated from each other- ‘*will’* and ‘*go’*: analytic behaviour.
- In case of present continuous tense expressed as ‘*going*’, the behaviour is synthetic- ‘*go*’ joined with ‘*ing*’.

The Phenomenon of Fusion/syncretism (1/2)

- Bound morphemes expressing grammaticality (number, tense etc.) or case relationships are overloaded, i.e., perform multiple roles
- One morpheme-one function is one end of the spectrum. The other end is small number of morphemes performing many morphological roles.
- Overloading of roles per morpheme is called ***syncretism***.
- In '*will go*', English is displaying syncretism (i.e., fusion), since *number* and *person* are

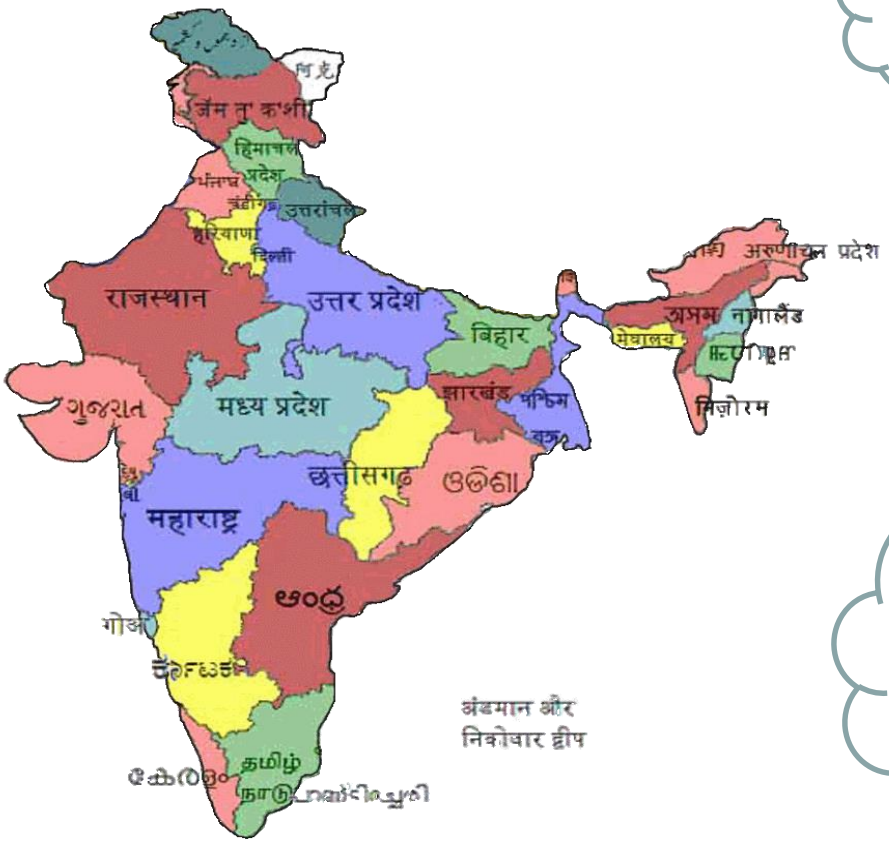
The Phenomenon of Fusion/syncretism (2/2)

- Overloading of roles per morpheme is called ***syncretism***.
- '*will go*': syncretism (i.e., fusion), since *number* and *person* are indeterminate here: "*I/we/you/he/she/they will do*".
- Hindi is much less syncretic than English- *jaaUmgaa* (first person, singular number, future tense of '*go*'), *jaaoge* (second person, singular number, future tense of '*go*'), *jaayegaa* (third person, singular number, future tense of '*go*').

Analytic-Synthetic Spectrum

Analytic

Synthetic



Hindi
Punjabi

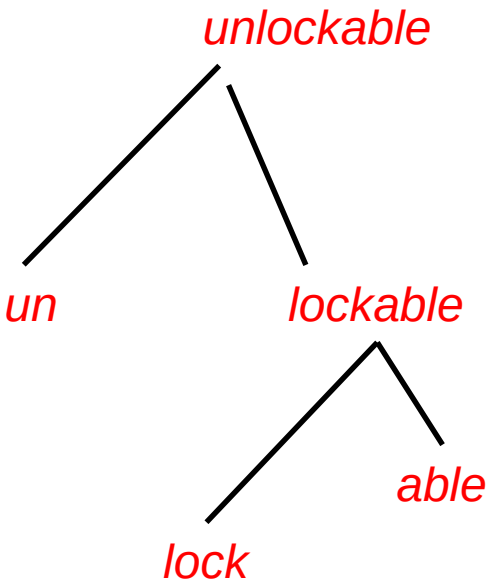
Gujarati
Marathi

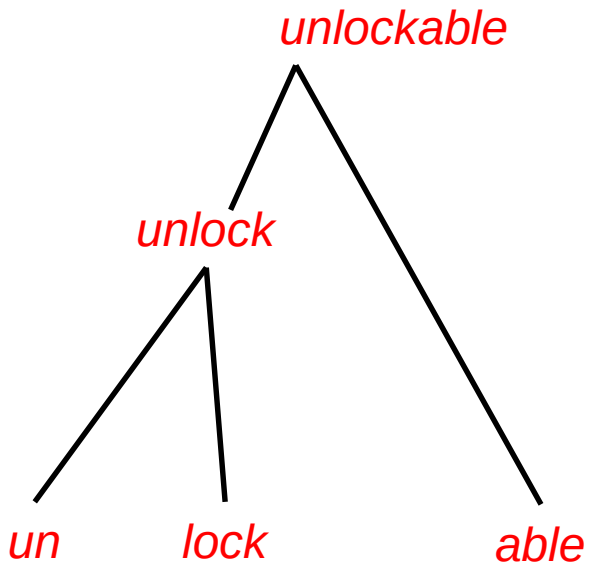
Tamil
Telugu
Kannad
Malayalam

Structural Typology Matrix

Stacking Of morphemes ↓ / Overloading of grammatical features →	**YES**	**NO**
YES	**Agglutinative-Fusional *(Turkish, Dravidian)***	**Agglutinative**
NO	**Isolating-Fusional *(English, Indo-European including Hindi)***	**Isolating *(Vietnamese, Chinese)***

Morphological Compositionality: Morphotactics of "Unlockable": two structures





something that cannot be locked ("*this gate is unlockable*"- open and cannot be locked)

Something that can be unlocked ("*this gate is unlockable*"- shut with a lock, but can be unlocked)

Language differ in morphological complexity (1/2)

- **Hindi-** *जॉन हर रोज स्कूल में बच्चों को रंगीन चाक से चित्र बनाना सिखाता है। jon har roj skool mein bachchon ko rangeen chaak se chitr banaana sikhaata hai |* (14 tokens; isolating behaviour)

- **Marathi-** *जॉन दररोज शाळेत मुलांना रंगीत खडूंनी चित्र काढायला शिकवतो| Jŏna dararōja śāḷēta mulānnā raṅgīta khaḍūnnī citra kāḍhāyalā śikavatō.* (9 tokens; agglutinating behaviour)

- **Bengali-** *জন প্রতিদিন স্কুলে বাচ্চাদের রঙিন চক দিয়ে আঁকা শেখান |* *Jana pratidina skulē bāccādēra raṅina caka diẏē āṁkā śēkhāna* (9 tokens)

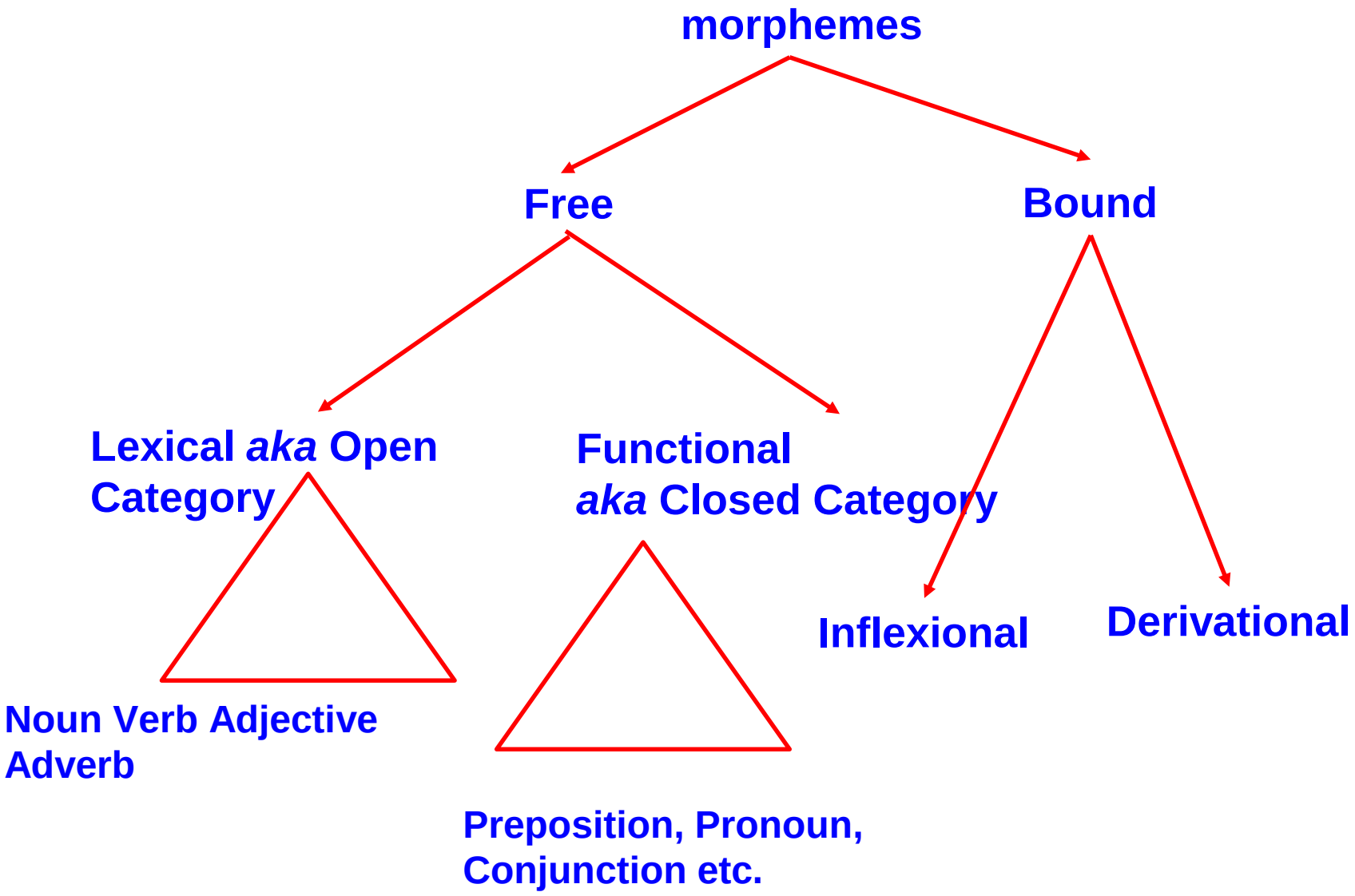
Basics of Morphology

Morphology

- The field of linguistics that examines internal structure of words and word formation rules

- Borrowing from Kenneth Pike's famous quote on Phonology:
  - "Morphemes ARE raw material, morphology cooks them"

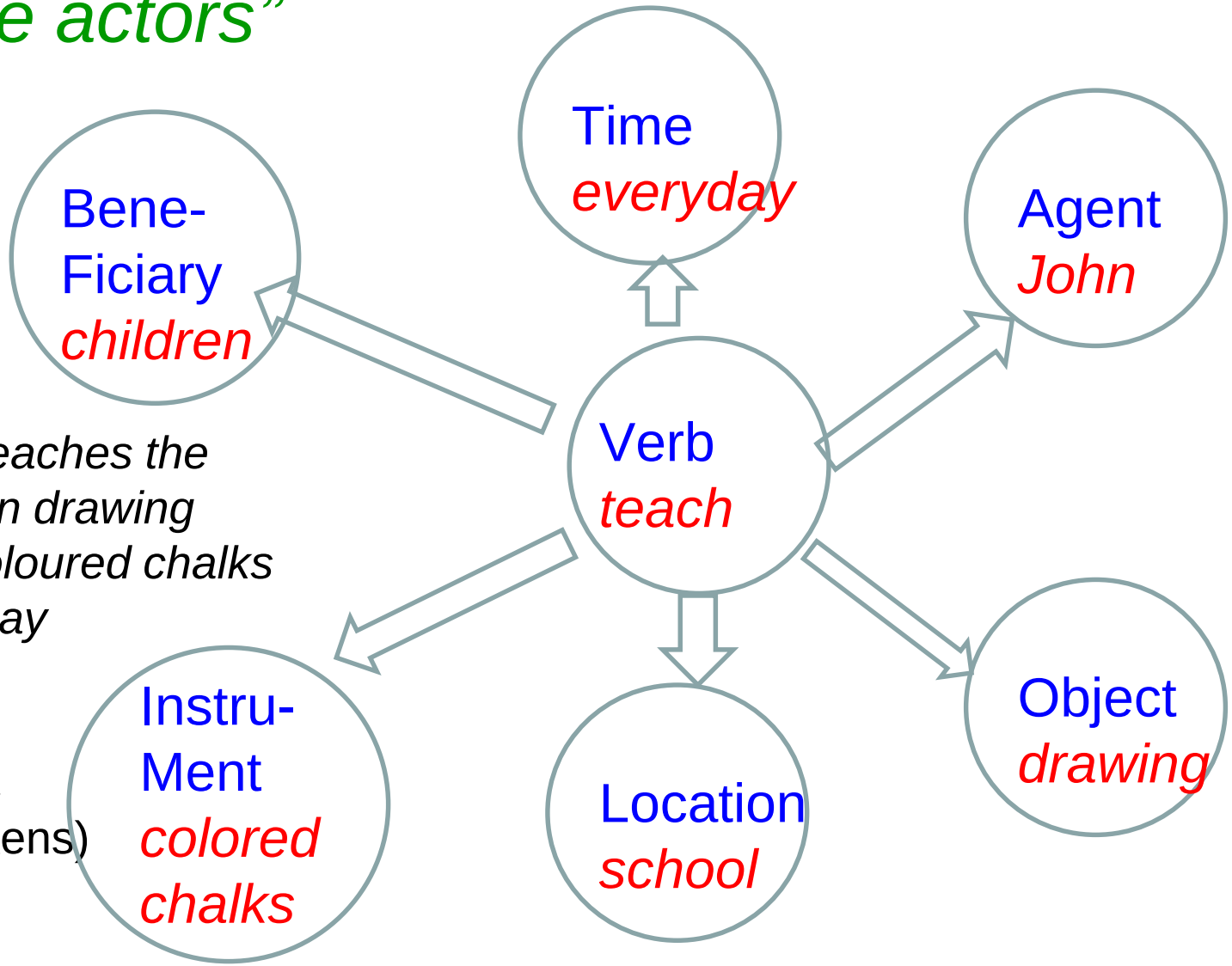
MORPHEME Ontology



Morphological Processes

- Inflection (*toy → toys*)
- Derivation (*describe → describable*)
- Compounding (*school-bus*)
- Portmanteau (*brunch → breakfast+lunch; smog →smoke+fog*)

Reason for morphology: #1: Case Roles: *"all the sentence is the stage and the words are the actors"*



John teaches the children drawing with coloured chalks everyday in the school.
(12 tokens)

Reason-2: Grammatical Features (exemplified with Hindi)

- Noun: inflects for number and case
  - लड़का /lʌr.kaː/ (s+nominative)
  - लड़के /lʌr.keː/ (pl+nominative or s+oblique)
  - लड़कों /lʌr.kõ/ (pl+oblique)

  (oblique means cases other than nominative)

- Verbs: inflects for gender (G), number (N), person (P), tense (T), aspect (A), modality (M) (GNPTAM)

GNPTAM with Hindi verb *"jaanaa" (to go)*

- G: *jaayegaa (he will go), jaayegii (she…)*
- N: *jaaeMge (they will go)*
- P: *jaayegaa (he will go), jaayoge (you will go), jaaUMgaa (I will go)*
- T: *jaataa_hEI (he goes), gayaa_thaa (he went), jaayegaa (he will go)*
- A: *gayaa (he has gone)*
- M: *jaaye (let him go), jaao (you go)*

Back to Discriminative Modeling

Modelling Equation

The probability that the tag at a position *i* in the word sequence

W: $\wedge w_0 w_1 w_2 \ldots w_{n-2} w_{n-1} w_n$. is *t* is given as

$$P(t_i = t \mid F_i) = \frac{e^{\sum_{j=1.k} \lambda_j f_{ij}}}{\sum_{t' \in S} e^{\sum_{j=1.k} \lambda_j f_{ij}(t')}}$$

where *S* is the set of tags. The sequence probability of a tag sequence *T* is as per equation (8), the product of $P(t_i|F_i)$, *i* varying over the positions.

Beam Search Based Decoding

- ^ *The brown foxes jumped .*
- Let us assume the following tags for the purpose of the discussion:
- D- determiner like ‘the’
- adjective like ‘brown’
- N- noun like ‘foxes’, ‘fence’
- V- verb like ‘jumped’
- Let the decoder start at the state ‘^’ which denotes start of the sentence.

Step-1

- The word '*the*' is encountered. First there are 4 next states possible corresponding to 4 tags, giving rise to 4 possible paths:
- ^ *D* -P_1
- ^ *A* -P_2
- ^ *N* -P_3
- ^ *V* -P_4

Commit to Beam Width

- Beam width is an integer which denotes how many of the possibilities should be kept open.
- Let us suppose that we decide the beam width is 2. This means that out of all the paths obtained so far we retain only the top 2 in terms of their probability scores.
- We will assume here that we get the actual linguistically viable sub-sequences as the top two choices. ‘The’ is a determiner and we get the two highest probability paths for “^ The” as P_1 and P_3.

to be continued….