# Wordnet, Multiword, Metaphor and UW

Pushpak Bhattacharyya

Department of Computer Science and Engineering

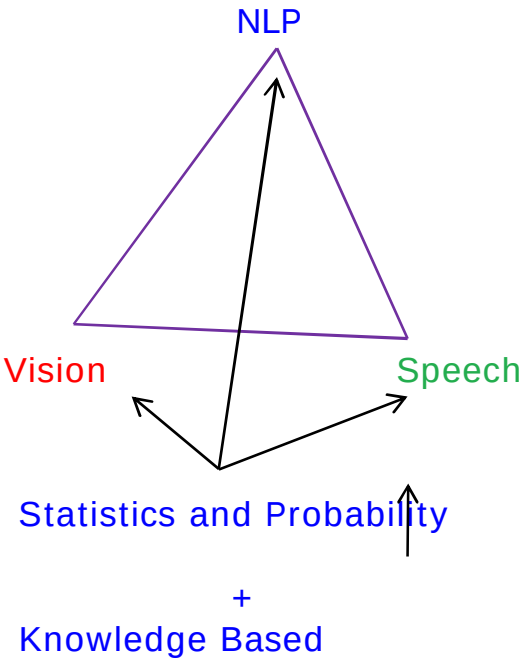
IIT Bombay

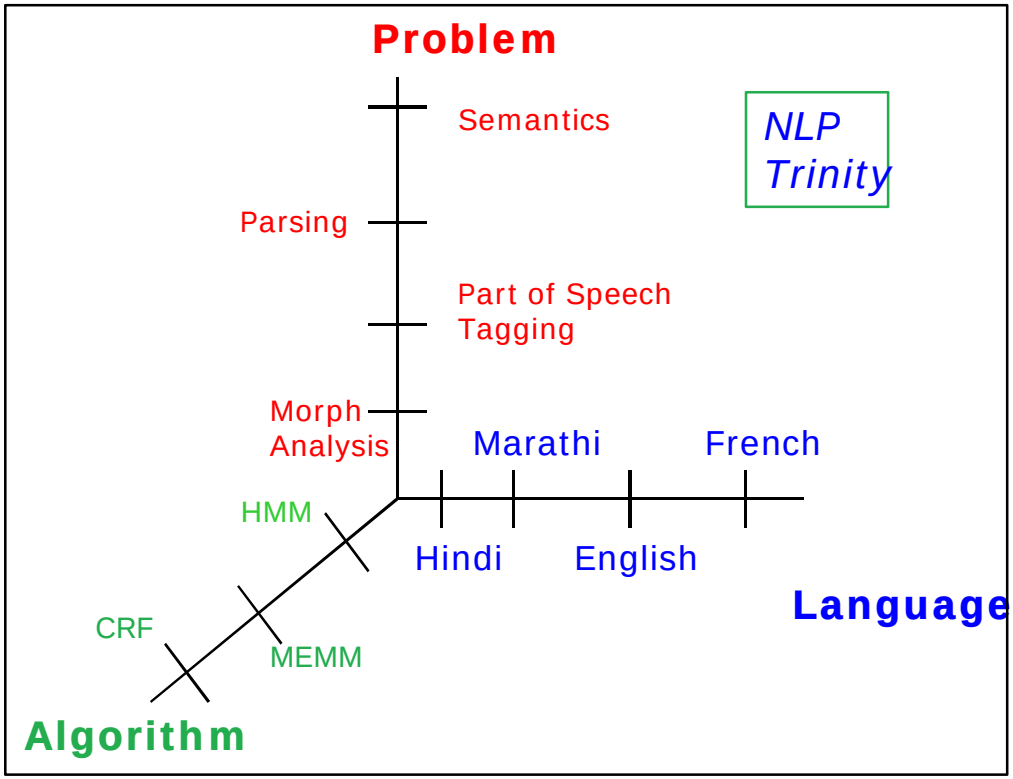
*COLING 2012 UNL Panel, IIT Bombay*

*15 December, 2012*

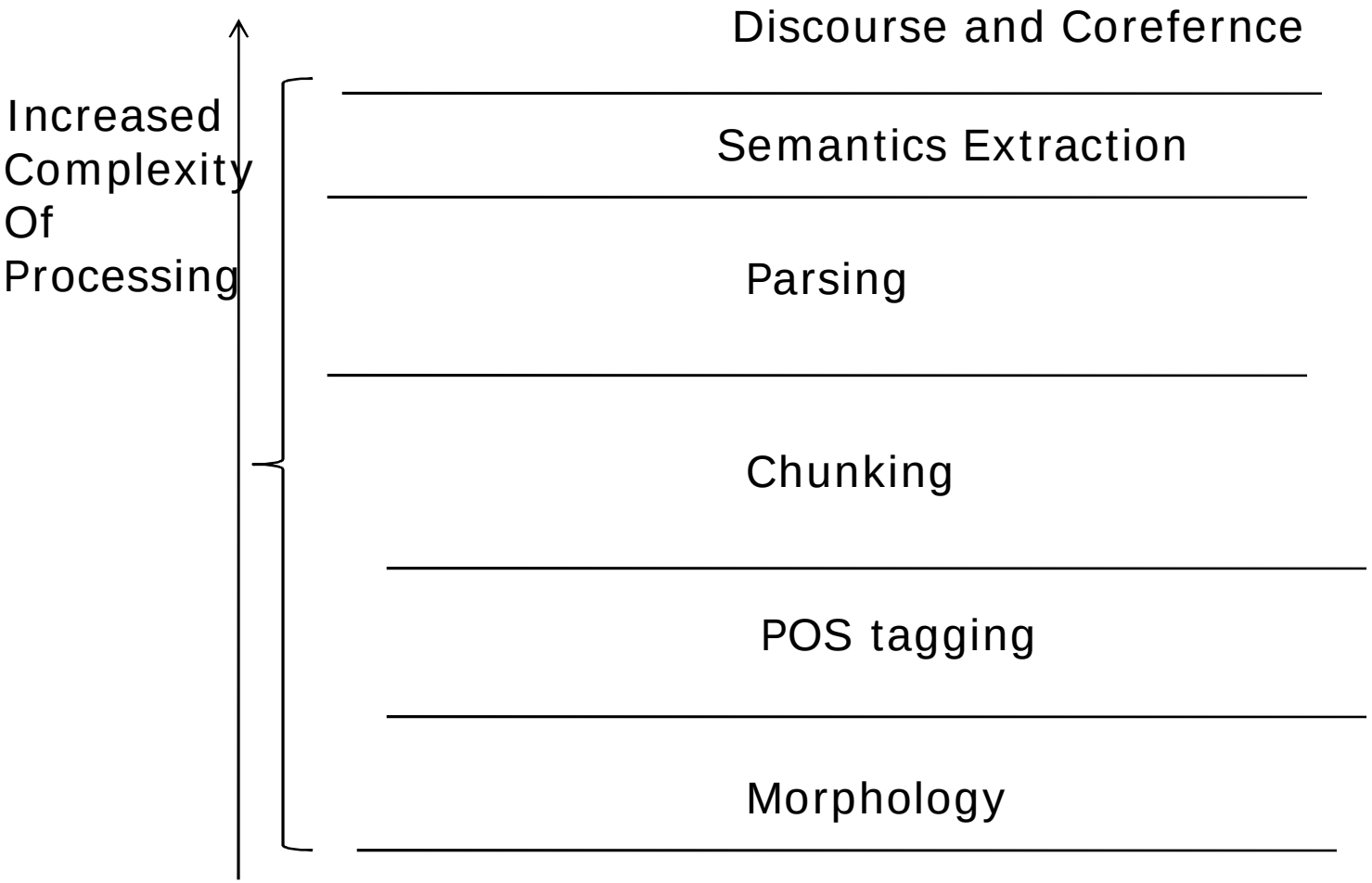
# Foundations

# Two pictures



# NLP Layer


Increased
Complexity
Of
Processing
Discourse and Corefernce
Semantics Extraction
Parsing
Chunking
POS tagging
Morphology

# Relational Semantics

| **Word Meanings** | **Word Forms** | | | | |
|---|---|---|---|---|---|
| | **$F_1$** | **$F_2$** | **$F_3$** | **…** | **$F_n$** |
| **$M_1$** | (*depend*) $E_{1,1}$ | (*bank*) $E_{1,2}$ | (rely) $E_{1,3}$ | | |
| **$M_2$** | | *(bank)* $E_{2,2}$ | | (*embankment*) $E_{2,…}$ | |
| **$M_3$** | | *(bank)* $E_{3,2}$ | $E_{3,3}$ | | |
| **…** | | | | … | |
| **$M_m$** | | | | | $E_{m,n}$ |

# Componential Semantics

- Consider *cat* and *tiger.* Decide on *componential attributes.*

| Furry | Carnivorous | Heavy | Domesticable |
|---|---|---|---|

- For *cat* (Y, Y, N, Y)
- For *tiger* (Y,Y,Y,N)

**Complete and correct Attributes are difficult to design.**

# Fundamental Design Question

- **Syntagmatic** *vs.* **Paradigmatic relations?**
- Psycholinguistics is the basis of the design.
- When we hear a word, many words come to our mind *by association*.
- For English, about half of the associated words are *syntagmatically related* and half are *paradignatically related.*
- For *cat*
  - *animal, mammal*- paradigmatic
  - *mew, purr, furry*- syntagmatic

# Coming to UW…

# Universal Word

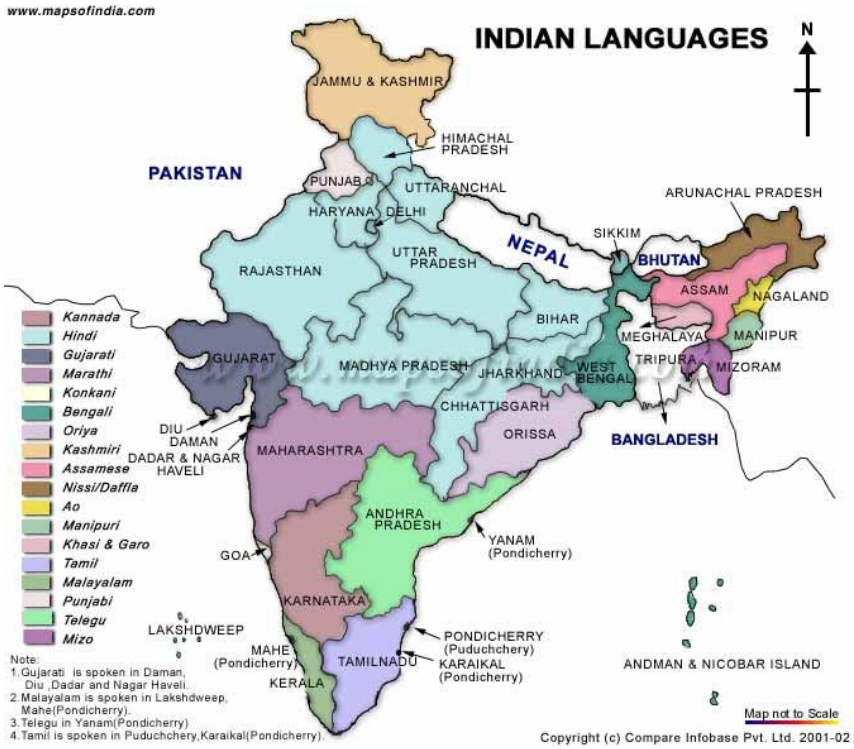
- The repository of Uws is supposed to be **Universal**
- Maybe the entities themselves are not!
- *Every concept expressed in every language should **find** a place in the UW dictionary*

# IITB's NLP effort and UW++

- Connect Indian languages to the other languages of the world through a pivot of interlingual lexemes, that will make machine translation easier among these languages.

# Indian Languages: a complex landscape

- Major streams
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Some languages are ranked within 20 in the world in terms of the populations speaking them
  - Hindi and Urdu: 5th (~500 milion)
  - Bangla: 7th (~300 million)
  - Marathi 14th (~70 million)




***TDIL program of DIT, Ministry of IT Launched large consortia projects on MT and IR***

# Some UW++ entries which are MWs

- Cabman
  - "cabman(icl>driver>thing,equ>taxidriver)" {n} "SOMEONE WHO DRIVES A TAXI FOR A LIVING" ""
- E [cabman] {CABMAN:AGENS,COUNT,STRONGCOUNT}
- F [chauffeur_de_taxi] {CAT(CATN),GNR(MAS)}

# Another multiword UW

- "counterbalance(icl>cancel>do, equ>counteract, agt>thing, obj>thing)” {v} "OPPOSE AND MITIGATE THE EFFECTS OF CONTRARY ACTIONS” "THIS WILL COUNTERACT THE FOOLISH ACTIONS OF MY COLLEAGUES"
- "counterbalance(icl>balance>be, equ>compensate, obj>thing,aoj>thing)” {v}"ADJUST FOR” "ENGINEERS WILL WORK TO CORRECT THE EFFECTS OR AIR RESISTANCE"
- "counterbalance(icl>contrast>do, equ>oppose, agt>thing, obj>thing)” {v} "OPPOSE WITH EQUAL WEIGHT OR FORCE"
- "counterbalance(icl>structure>thing, equ>balance)” {n} "EQUALITY OF DISTRIBUTION"
- "counterbalance(icl>weight>thing, equ>counterweight)” {n} "A WEIGHT THAT BALANCES ANOTHER WEIGHT"

# UW dictionary is a linked structure like the wordnet

- "waddle(icl>walk>do,equ>toddle,agt>thing)" {v} "WALK UNSTEADILY" "SMALL CHILDREN TODDLE"

- toddle, coggle, totter, dodder, paddle, waddle -- (walk unsteadily; "small children toddle")

  => walk -- (use one's feet to advance; advance by steps; "Walk, don't run!")

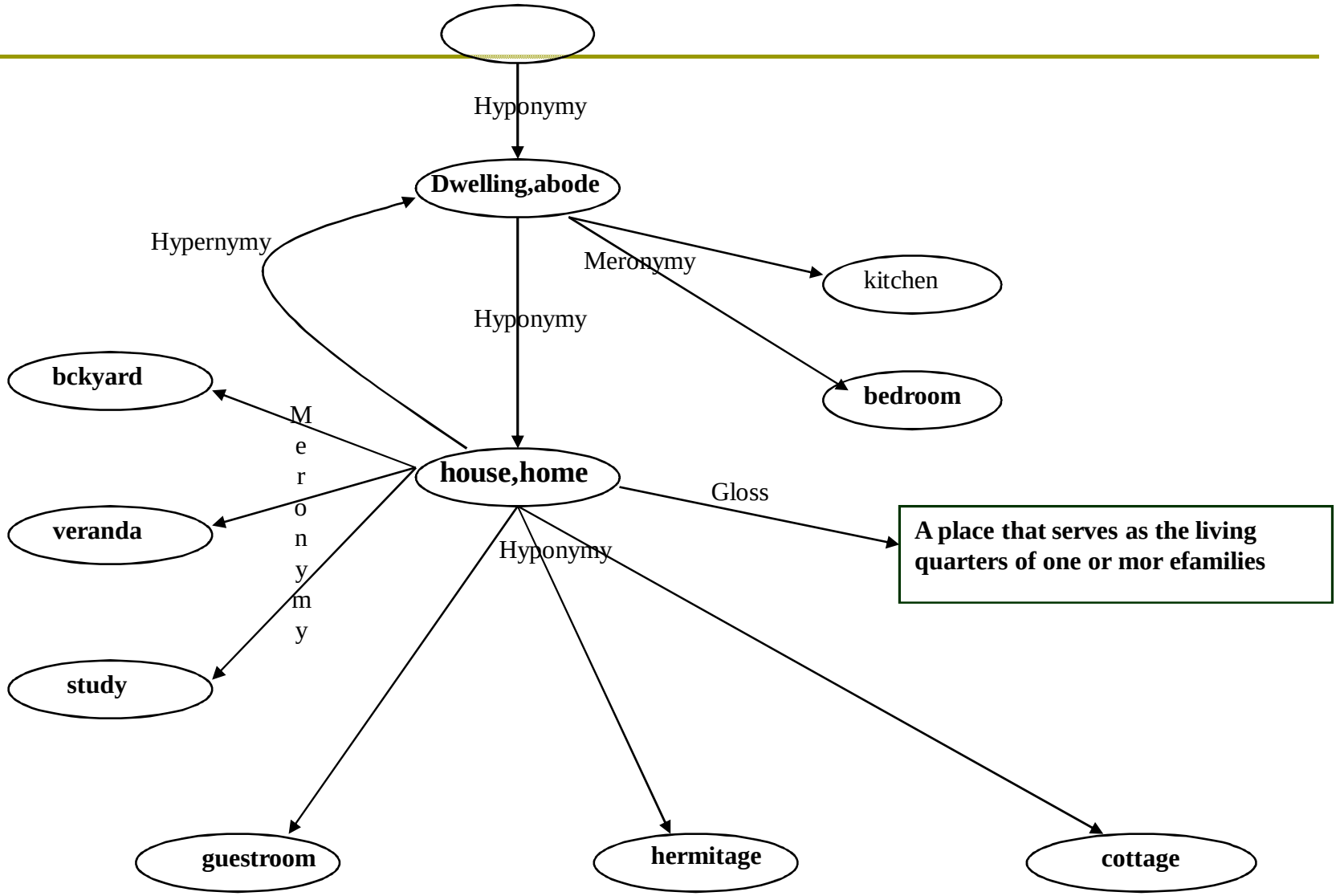
  => travel, go, move, locomote -- (change location; move, travel, or proceed; "How fast does your new car go?")

# Lexical and Semantic relations in wordnet

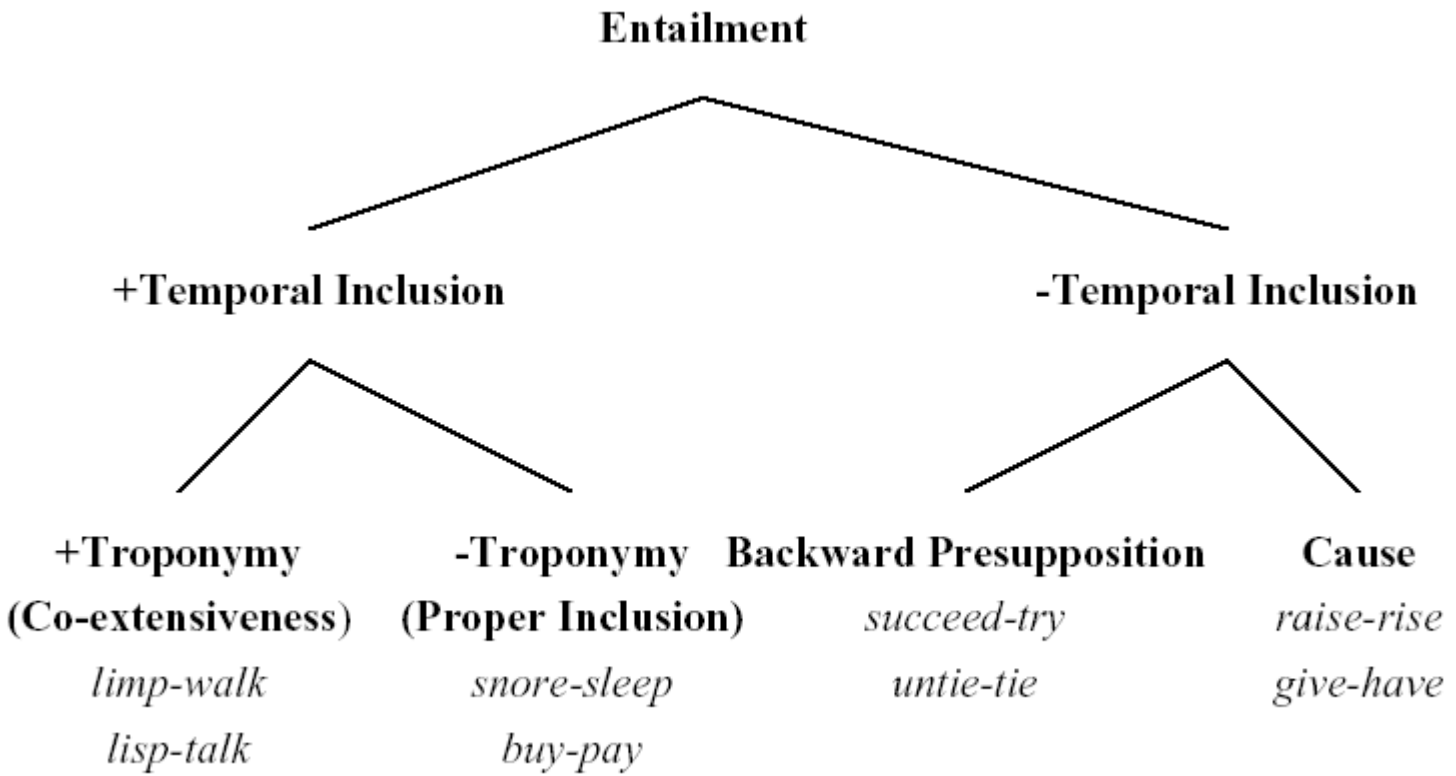
1. Synonymy
2. Hypernymy / Hyponymy
3. Antonymy
4. Meronymy / Holonymy
5. Gradation
6. Entailment
7. Troponymy

1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

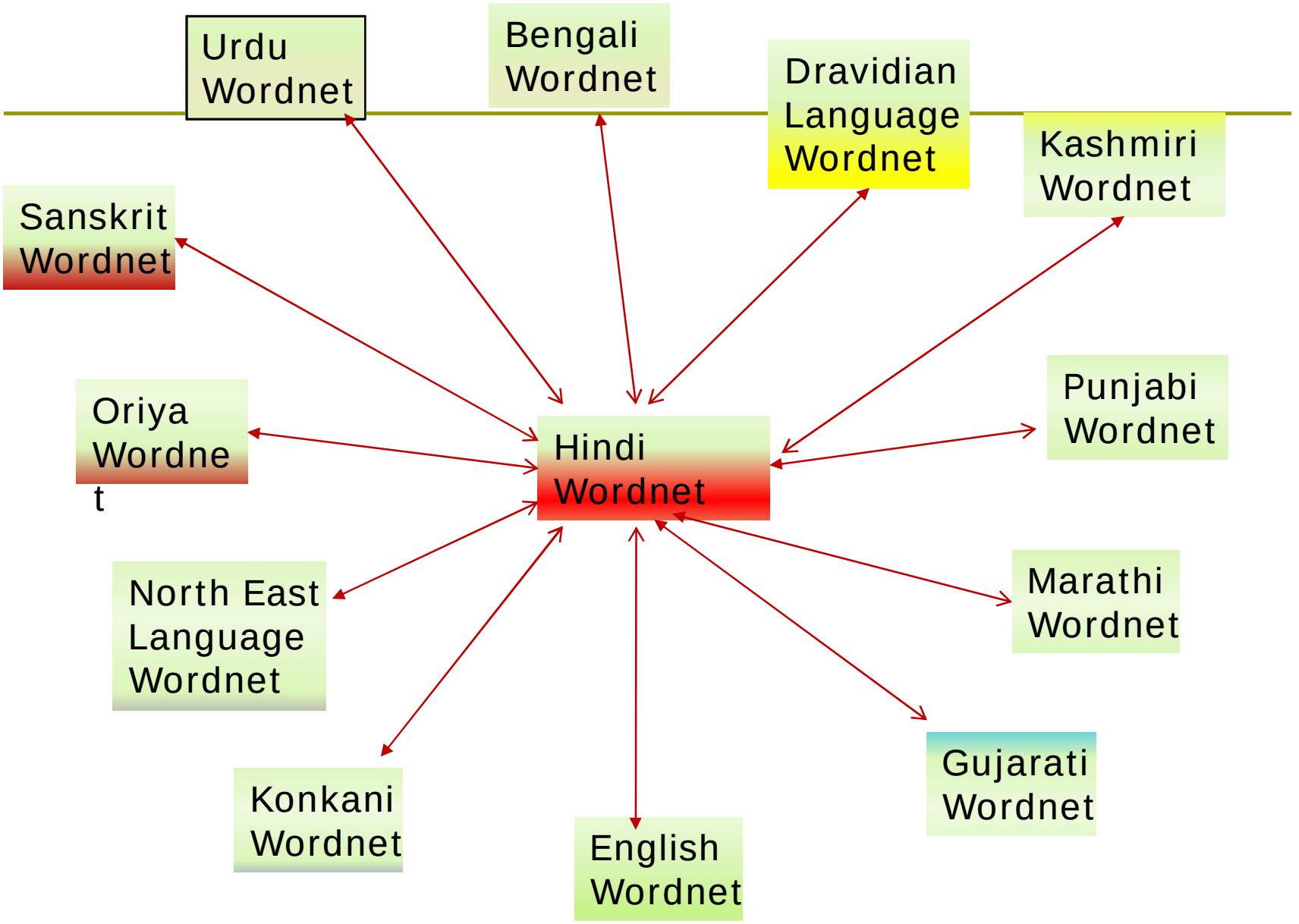
# WordNet Sub-Graph



# Verbs in wordnet



# INDOWORDNET



# Categories of Synsets (2/2)

- **Language specific**: Synsets which are unique to a language (*e.g. Bihu* in Assamese language)

- **Rare**: Synsets which express technical terms (*e.g. ngram*).

- **Synthesized:** Synsets created in the language due to influence of another language (*e.g. Pizza*).

# Need for categorization

- To bring systematicity in the way the wordnet synsets are linked
  - Universal→Pan Indian→Language Family→Language→Synthesised→Rare
- All members have finished the Universal and Pan Indian synsets

# Categorization methodology

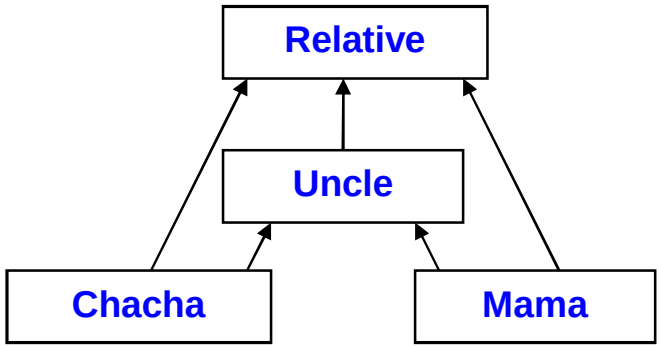
- 34378 Hindi synsets were sent to all Indo-wordnet groups in the tool, in which they had these options to categorize:
  - **Yes**
  - **No**

- **Universal synsets:-** The synsets which were categorized Yes and also have equivalent English words or synsets.
- **Pan-Indian :-** The synsets which were categorized Yes and did not have equivalent English words or synsets.

# Expansion approach: linking is a subtle and difficult process

- To link or not to link
- While linking:
  - face lexical and semantic chasms
  - Syntactic divergences in the example sentences
    - Change of POS
    - Copula drop (Hindi→Bangla)

# Linking kinship relations and fine grained concepts



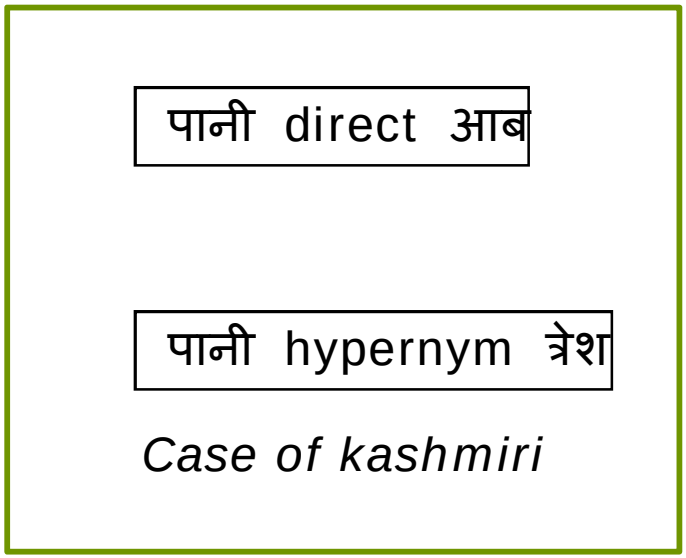

पानी direct आब

पानी hypernym त्रेश

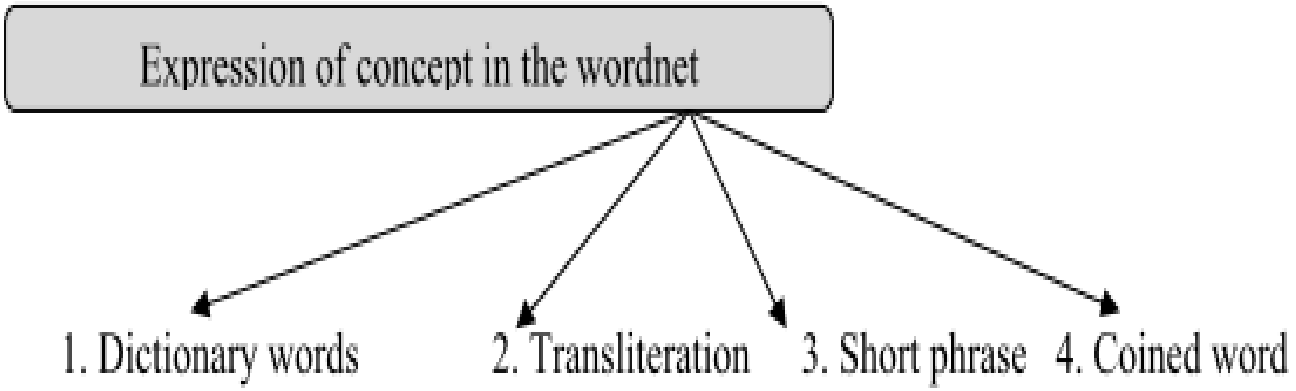
*Case of kashmiri*

# Important decision

- TWO kinds of linkages
  - Direct
  - Hypernymy



# How to express a concept not present in the language?



# Transliteration: often employed

- Synset ID : 39 POS : adjective Synonyms : सनाथ, (*sanaatha*)
- Gloss : जिसका कोई पालन-पोषण या देखभाल करने वाला हो (opposite of *orphan*)
- Example statement : "सनाथ बालकों को अनाथ बालकों की मदद करनी चाहिए *(children who are looked after should help the orphans)*/ साधक प्रभु का हो जाने पर अनाथ नहीं रहता, सनाथ हो जाता है"

- Transliterated and adopted by Bangla and Gujarati

# Short phrase: often employed

| No. | Hindi Words | Concept | Bengali Equivalents |
|---|---|---|---|
| 1 | रगड़ाई, घिसाई | Wage for rubbing something | রগড়ানোর মজুরী |
| 2 | छिड़काई | Wage for scattering something | ছড়ানোর মজুরী |
| 3 | कटवाई | Wage for cutting something | কাটানোর মজুরী |

**Bangla**

| Hindi | Urdu |
|---|---|
| अमांगलिक,अमाङ्गलिक: | بدبختی،بدقسمتی،بدنصیبی |

**Urdu (meaning *Inauspicious)***

In the above words the prefix بد *(bad)* has been prefixed for the Hindi prefix अ.

# Linking synsets across languages: Influence on Hindi Wordnet

Hindi wordnet has to add new synsets to accommodate language specific concepts, e.g., in Gujarati

**ભૈરવજપ (bhairav jap)**

**ID :: 103040**

**CAT :: NOUN**

**CONCEPT :: मोक्ष के लिए जप करते हु एपर्वत पर से अपने आप को गिराना (Taking God's name and throwing oneself from atop a mountain to attain liberation)**

**EXAMPLE :: गिरनार के शिखर पर से यात्रिक भैरवजप करते थे एसा माना जाता है। (it is thought that pilgrms used to do bhairav jap atop Girnar mountain)**

**SYNSET-HINDI :: भैरवजप**

# Multiwords

# MWs can be long

Long Expressions with variable relationships

*Colon Cancer Tumor Suppressor Protein*

Head: Protein

**Mod (protein-5, suppressor-4); protein causing suppression**

**Mod (suppressor-4, tumor-3);**
**suppressor *causing* tumor (*)**
**suppressor */* suppressing *of* tumor**

**Mod (tumor-3, cancer-2); tumor *caused-by* cancer**

**Mod(cancer-2, colon-1); cancer *of* colon**

# Necessary and Sufficient Conditions for MWness

- Necessary Condition
  - Word sequence separated by space/delimiter
- Sufficient Conditions
  - Non-compositionality of meaning
  - Fixity of expression
    - In lexical items
    - In structure and order

# Examples – Necessary condition

- Non-MWE example:
  - Marathi: सरकार हक्काबक्का झाले
  - Roman: sarakAra HakkAbakkA JZAle
  - Meaning: government was surprised
- MWE example:
  - Hindi: गरीब नवाज़
  - Roman: garIba navAjZa
  - Meaning: who nourishes poor

# Examples - Sufficient conditions ( Non-compositionality of meaning)

- Konkani: पोटांत चाबता
- Roman: poTAMta cAbatA (literally, *biting in the stomach*)
- Meaning: to feel jealous

- Telugu: చెట్టు కిందికి ప్లీడరు
- Roman: ceVttu kiMda pLIdaru (literally, *a lawyer sitting under the tree)*
- Meaning: an idle person

- Bangla: মাটির মানুষ
- Roman: mAtira mAnuSa
- Meaning: a simple person/son of the soil

# Examples – Sufficient conditions (Fixity of expression)

## In lexical items

- Hindi
  - *usane muJe gAlI dI (he abused me)*
  - **usane muJe galI pradAna kI*
- Bangla
  - *jabajjIbana karadaMda (life imprisonment)*
  - **jIbanabhara karadaMda*
  - **jabajjIbana jela*
- English (1)
  - *life imprisonment*
  - **lifelong imprisonment*
- English (2)
  - *Many thanks*
  - **Plenty thanks*

# Examples – Sufficient conditions (In structure and order)

- English example
  - *kicked the bucket (died)*
  - *the bucket was kicked*

    *(not passivizable in the sense of dying)*
- Hindi example
  - *उम्र क़ैद*
  - *umra kEda (life imprisonment)*
  - *umra bhara kEda*

# Characterization of IL-MWs

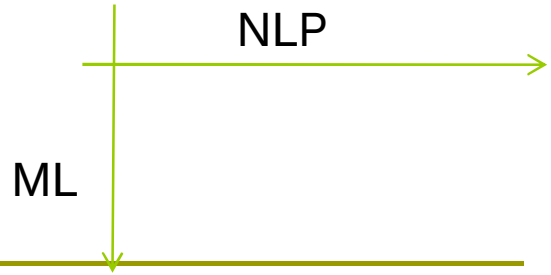
# Reduplicative MWs

- Complete
  - Onomatopoeic *(gutar gutar (Hindi)* meaning *sound made by pigeons*)
  - Non-Onomatopoeic (*ghar ghar (Hindi)* meaning *in every house*)
- Partial
  - With echo words (*pani vani (H)* meaning *water etc., bai tai (Bangla)* meaning *book etc.)*
  - With words of different origin (*pran thawai (Manipuri)* meaning *soul; sena lanmi (Manipuri)* meaning *army):* both composed of Sanskrit and Manipuri
  - With meaningless words (balancing compounds) (*irugu poVrugu (Telugu)* meaning *neighbours)*

# Non-Reduplicative MWs

- Synonyms (*ghar baAdI (Bangla)* meaning *houses/homes)*
- Antonym (*jannat jahannum (Urdu)* meaning *heaven and hell)*
- Complex predicates
  - Conjunct verbs (*kiTappil* 'in state of lying' + *pooTu* > *kiTappil pooTu* 'keep something pending‘ (Tamil)*)*
  - Compound verbs (*faao khalam (Bodo)* meaning *to finish acting on a task*)

# MW task (NLP + ML)



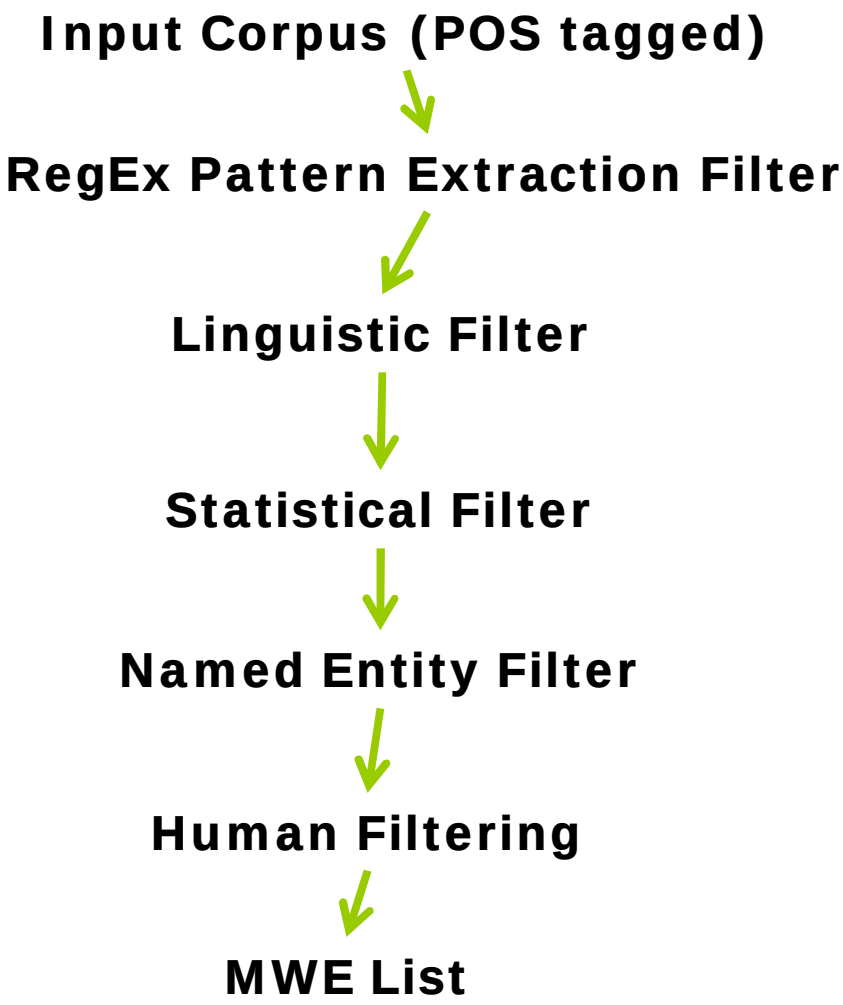

| | ***String + Morph*** | ***POS*** | ***POS+ WN*** | ***POS + List*** | ***Chu nk-ing*** | ***Parsing*** |
|---|---|---|---|---|---|---|
| ***Rules*** | Onomaetopic Reduplication *(tik tik, chham chham)* | Non-Onomaetopic Reduplication *(ghar ghar)* | Non-redup (Syn, Anto, Hypo) *(raat din, dhan doulat)* | | | Non-contiguous something |
| ***Statistical*** | | Collections or fixed expressions *(many thanks)* | | Conjunct verb (verbalizer list), Compund verb (verctor verb list) *(salaha dena, has uthama)* | | Non-contiguous Complex Predicate |

***Idioms will be list morph + look up***

# MWE Extraction Engine: pipeline architecture

- Developed at IIT Bombay to extract Multiwords from input corpus
- Combination of filters
- MWE list produced after passing the corpus through the pipeline

# MWE Pipeline



# Metonymy

# Metonymy

- Associated with *Metaphors* which are epitomes of semantics
- Oxford Advanced Learners Dictionary definition: “The use of a word or phrase to mean something different from the literal meaning”

# Insight from Sanskritic Tradition

- Power of a word
  - Abhidha, Lakshana, Vyanjana
- Meaning of **Hall:**
  - *The hall is packed (avidha)*
  - *The hall burst into laughing (lakshana)*
  - *The Hall is full (*unsaid*: and so we cannot enter) (vyanjana)*
- How will **hall** be represented in these three cases, in the UW dictionary?

# Metaphors in Indian Tradition

- *upamana* and *upameya*
  - Former: object being compared
  - Latter: object being compared with
  - *Richard the Lion* (Richard: *upameya;* Lion: *upamana)*

# *Upamana, rupak, atishayokti*

- *upamana*: Explicit comparison
  - *King Richard was like a lion leading the crusaders*
- *rupak*: Implicit comparison
  - *King Richard was a lion leading the crusaders*
- *Atishayokti (exaggeration)*: upamana and upameya dropped
  - *King Richard led the crusaders from the front. The lion was everywhere in the battlefield.*

# Modern study (1956 onwards, Richards et. al.)

- Three constituents of metaphor
  - *Vehicle* (items used metaphorically)
  - *Tenor* (the metaphorical meaning of the former)
  - *Ground* (the basis for metaphorical extension)
- *"The foot of the mountain"*
  - Vehicle: :foot"
  - Tenor: "lower portion"
  - Ground: "spatial parallel between the relationship between the foot to the human body and the lower portion of the mountain with the rest of the mountain"

# Interaction of semantic fields (*Haas*)

- Core vs. peripheral semantic fields
- Interaction of two words in metonymic relation brings in new semantic fields with selective inclusion of features
- *Leg of a table*
  - Does not *stretch* or *move*
  - Does *stand* and *support*

# Lakoff's (1987) contribution

- Source Domain
- Target Domain
- Mapping Relations

# Mapping Relations: ontological correspondences

- *Anger is heat of fluid in container*

| Heat | Anger |
|---|---|
| (i) Container | Body |
| (ii) Agitation of fluid | Agitation of mind |
| (iii) Limit of resistence | Limit of ability to suppress |
| (iv) Explosion | Loss of control |
| | |

# Image Schemas

- Categories: Container Contained
- Quantity
  - More is up, less is down: *Outputs rose dramatically; accidents rates were lower*
  - Linear scales and paths: *Ram is by far the best performer*
- Time
  - Stationary event: *we are coming to exam time*
  - Stationary observer: *weeks rush by*
- Causation: *desperation drove her to extreme steps*

# Patterns of Metonymy

- Container for contained
  - *The kettle boiled* (water)
- Possessor for possessed/attribute
  - *Where are you parked?* (car)
- Represented entity for representative
  - The government will announce new targets
- Whole for part
  - *I am going to fill up the car with petrol*

# Patterns of Metonymy *(contd)*

- Part for whole
  - *I noticed several new faces in the class*
- Place for institution
  - *Lalbaug witnessed the largest Ganapati*

*Question: Can you have part-part metonymy*

# Feature sharing not necessary

- In a restaurant:
  - *Jalebii ko abhi dudh chaiye*

    (‘the jalebi (a sweet) now wants milk’)
    - no feature sharing
  - *The elephant now wants some coffee* (feature sharing)
  - (a fat man desiring coffee)

# Proverbs

- Describes a specific event or state of affairs which is applicable metaphorically to a range of events or states of affairs provided they have the same or sufficiently similar image-schematic structure

# Investigation into Sanskritic traditions

- Rich work of *smAsa* and their types
- Concept of *sAmarthya*
- When can adjacent words *combine* to give a single meaning?
- Example:
  - *krishnena bhramarena daMshitavati radha rorudyamati cha (bitten by the black bee Radha is crying)*
  - *krishabhramarena daMshitavati radha rorudyamati cha (bitten by the black bee Radha is crying)*
  - Helped by the same *subanta* (declension)
  - But modern descendents of Sanskrit have very little agreement between adjective and the qualified noun

# Conclusions (1/2)

- To ensure coverage, Uws need to represent MWs and metaphors
- More precision- if possible- needed in the **theory of uws**
  - *sensational(icl>adj,icl>good); two parents ??*
- Such a theory is needed, even if limited
- Can specify exceptions (like Panini)

# Conclusions (2/2)

- IMP: not all words in the sentence corresponds to a UW (but an attribute; e.g., *she seems disturbed; seems* should go as attribute)

- Named Entities (not covered) need to be
  - Detected only once
  - Stored for the future
  - Disambiguation needed (*Washington voted Washington to power*)
  - Very closely linked with coreference resolution

Thank You

http://www.cse.iitb.ac.in/~pb

http://www.cfilt.iitb.ac.in