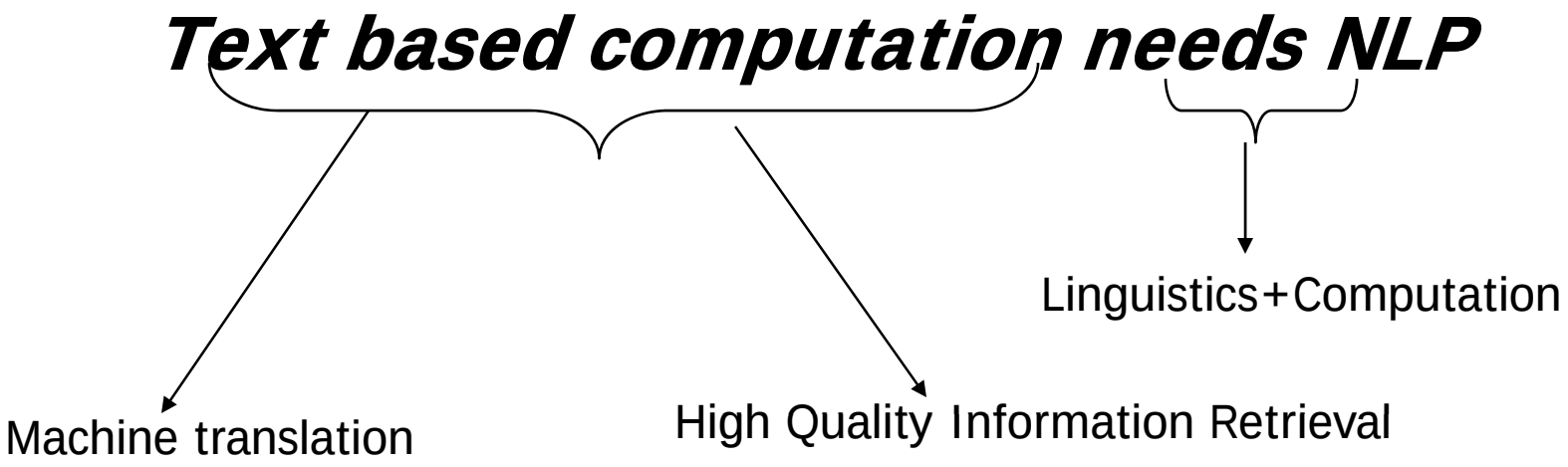
CS344: Introduction to Artificial Intelligence

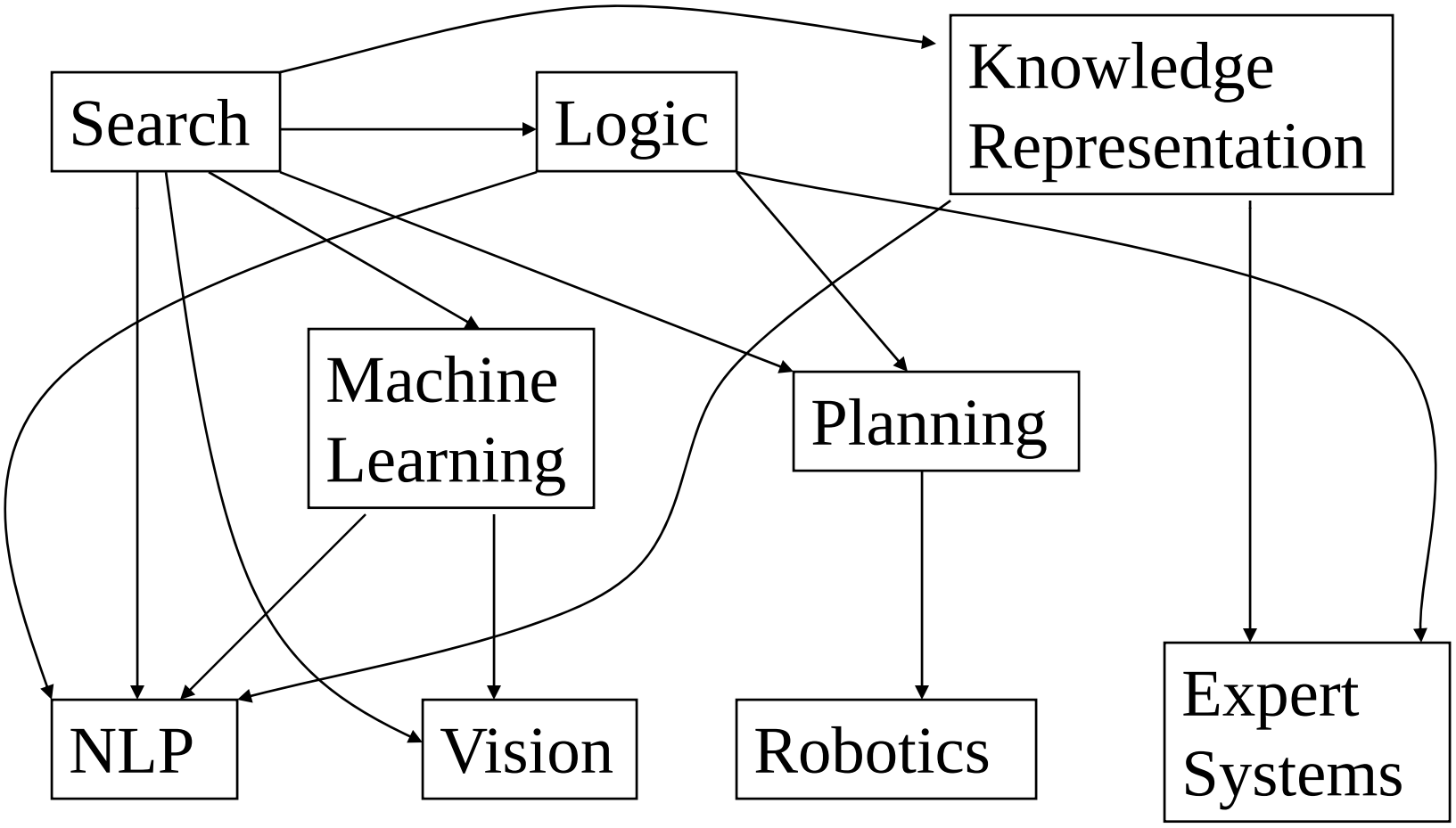
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 18-19– Natural Language Processing (ambiguities; Machine Learning and NLP)

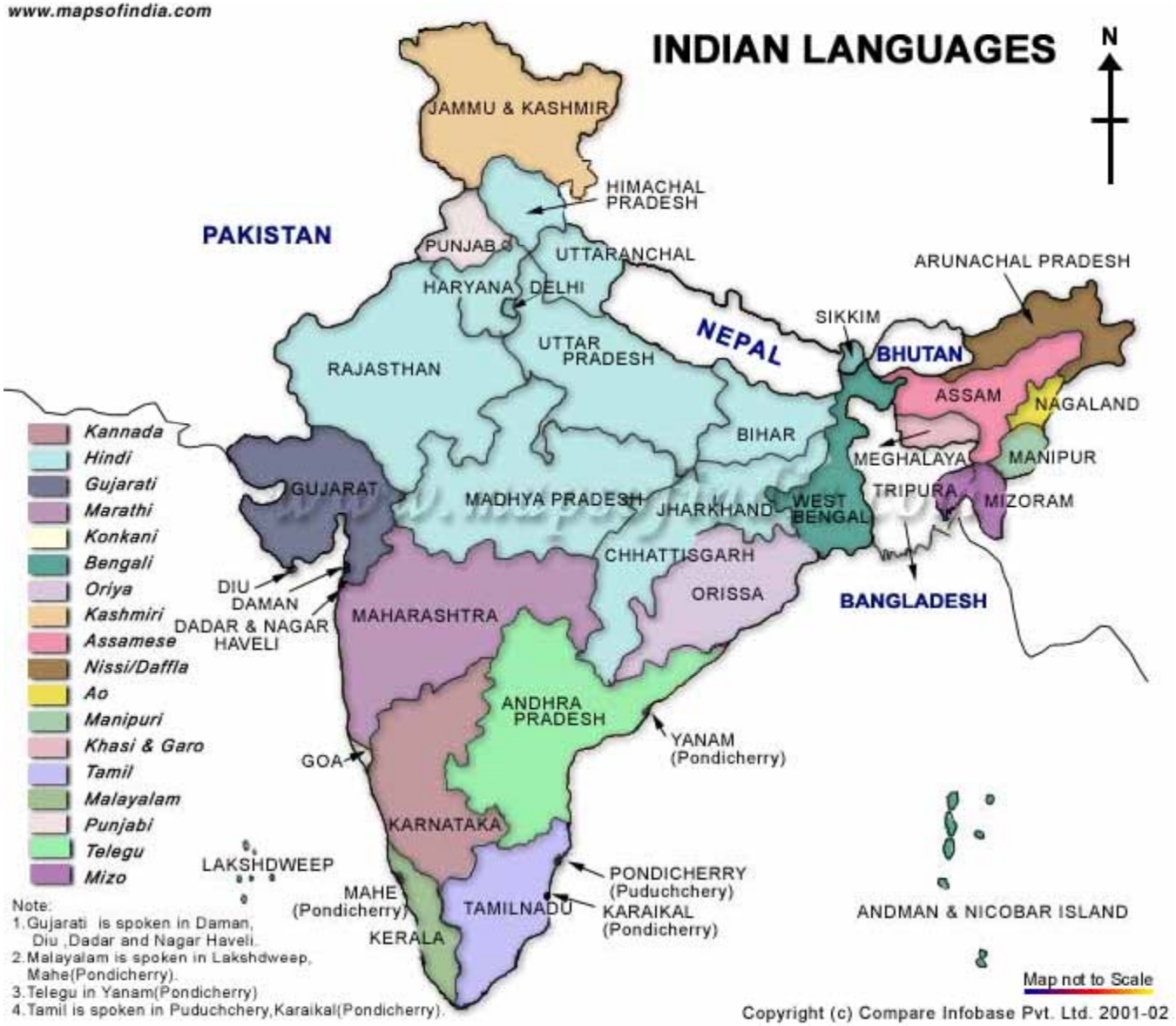
Importance of NLP



Perpectivising NLP: Areas of AI and their inter-dependencies



AI is the forcing function for Computer Science, and NLP of AI


www.mapsofindia.com
INDIAN LANGUAGES
N
JAMMU & KASHMIR
HIMACHAL PRADESH
PAKISTAN
PUNJAB
UTTARANCHAL
HARYANA
DELHI
NEPAL
SIKKIM
ARUNACHAL PRADESH
BHUTAN
ASSAM
NAGALAND
UTTAR PRADESH
RAJASTHAN
BIHAR
MEGHALAYA
MANIPUR
TRIPURA
MIZORAM
GUJARAT
MADHYA PRADESH
JHARKHAND
WEST BENGAL
CHHATTISGARH
ORISSA
BANGLADESH
DIU
DAMAN
DADAR & NAGAR HAVELI
MAHARASHTRA
ANDHRA PRADESH
YANAM (Pondicherry)
GOA
KARNATAKA
LAKSHDWEEP
MAHE (Pondicherry)
TAMILNADU
PONDICHERRY (Puduchchery)
KARAIKAL (Pondicherry)
KERALA
ANDMAN & NICOBAR ISLAND
Kannada
Hindi
Gujarati
Marathi
Konkani
Bengali
Oriya
Kashmiri
Assamese
Nissi/Daffla
Ao
Manipuri
Khasi & Garo
Tamil
Malayalam
Punjabi
Telegu
Mizo
Note:
1. Gujarati is spoken in Daman, Diu ,Dadar and Nagar Haveli.
2. Malayalam is spoken in Lakshdweep, Mahe(Pondicherry).
3. Telegu in Yanam(Pondicherry)
4. Tamil is spoken in Puduchchery,Karaikal(Pondicherry).
Map not to Scale
Copyright (c) Compare Infobase Pvt. Ltd. 2001-02

Languages and the speaker population

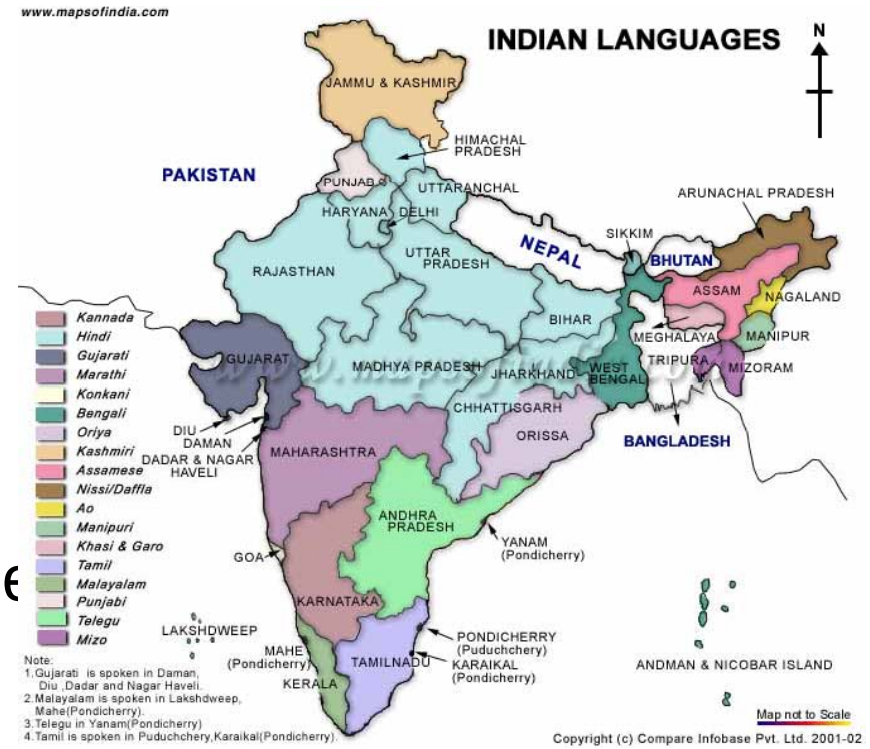
Language	**Population (2001 census; rounded to most significant digit)**
Hindi	450, 000, 000
Marathi	72, 000, 000
Konkani	7, 000, 000
Sanskrit	6000
Nepali	13, 000, 000

Languages and the speaker population (contd.)

Language	**Population (2001 census; rounded to most significant digit)**
Kashmiri	5, 000, 000
Assamese	13, 000, 000
Tamil	60, 000, 000
Malayalam	33, 000, 000
Bodo	1, 000, 000
Manipuri	1, 000, 000

Great Linguistic Diversity

- 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

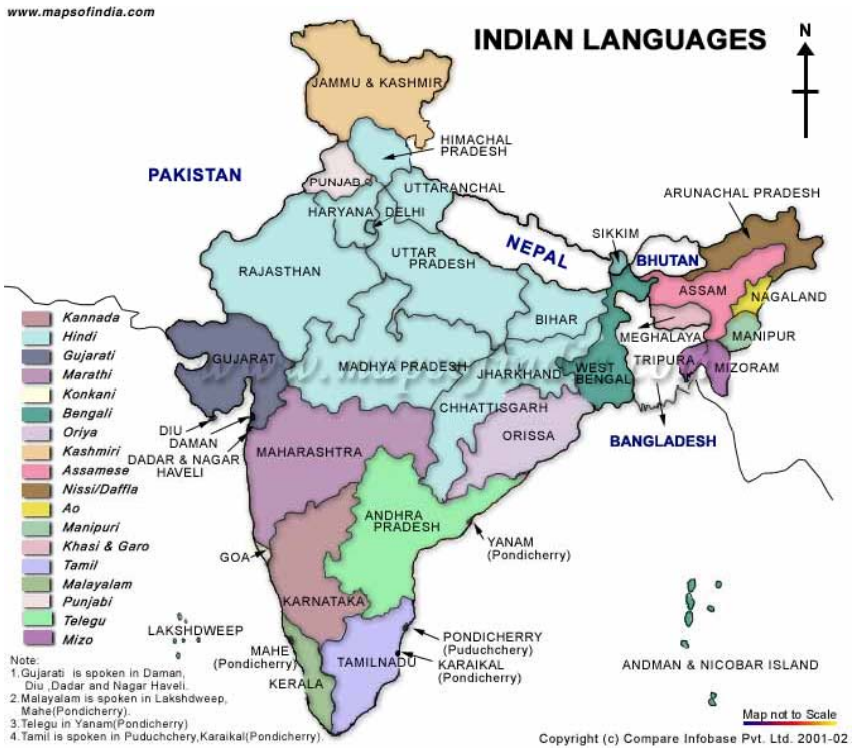


Interesting “mixed-race” languages

- **Marathi and Oriya**: confluence of *Indo Aryan* and *Dravidian* families
- **Urdu:** structure from Indo Aryan (Hindi), vocabulary from Persian and Semitic (Arabic)
  - आज मेरी परीक्षा है (aaj merii pariikshaa hai) {today I have my examination}
  - आज मेरा इम्तहान है (aaj meraa imtahaan hai)

3 Language Formula

- Every state has to implement
  - *Hindi*
  - *The state language (Marathi, Gujarathi, Bengali etc.)*
  - *English*
- Big time translation requirement, *e.g.*,during the financial year ends

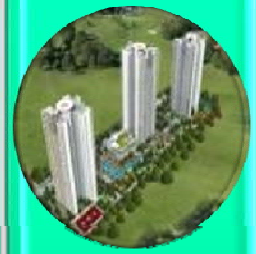


Multilingual Information Access needed for large GoI sector

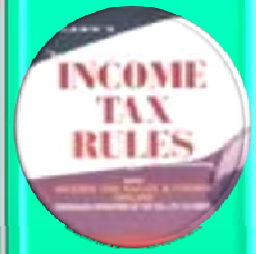
Provide one-stop access and insight into information related to key Government bodies and execution areas

Enable citizens exercise their fundamental rights and duties

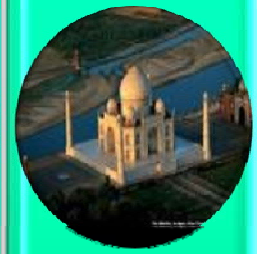
Science



Housing

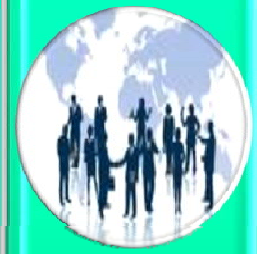


Taxes

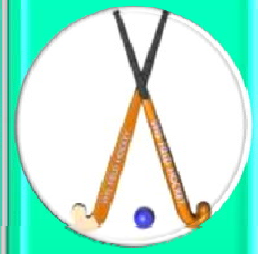


Travel & Tourism

Banking & Insurance



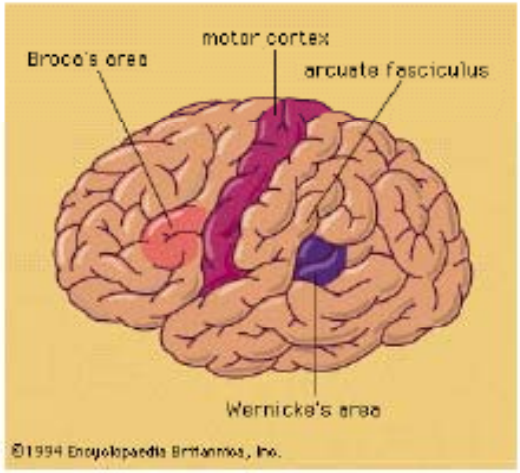
International



Sports

Need for NLP

- Machine Translation
- Information Retrieval and Extraction with NLP
  - Better precision and recall
- Summarization
- Question Answering
- Cross Lingual Search (very relevant for India)
- Intelligent interfaces (to Robots, Databases)
- Combined image and text based search
- Automatic Humour analysis and generation
- Last but not the least, window into human mind; *language and brain*



Broca's area: Region located anteriorly in the left hemisphere in the left frontal lobe operculum. It is responsible for production of words and sentences. This area is named after Paul Broca (in 1861).

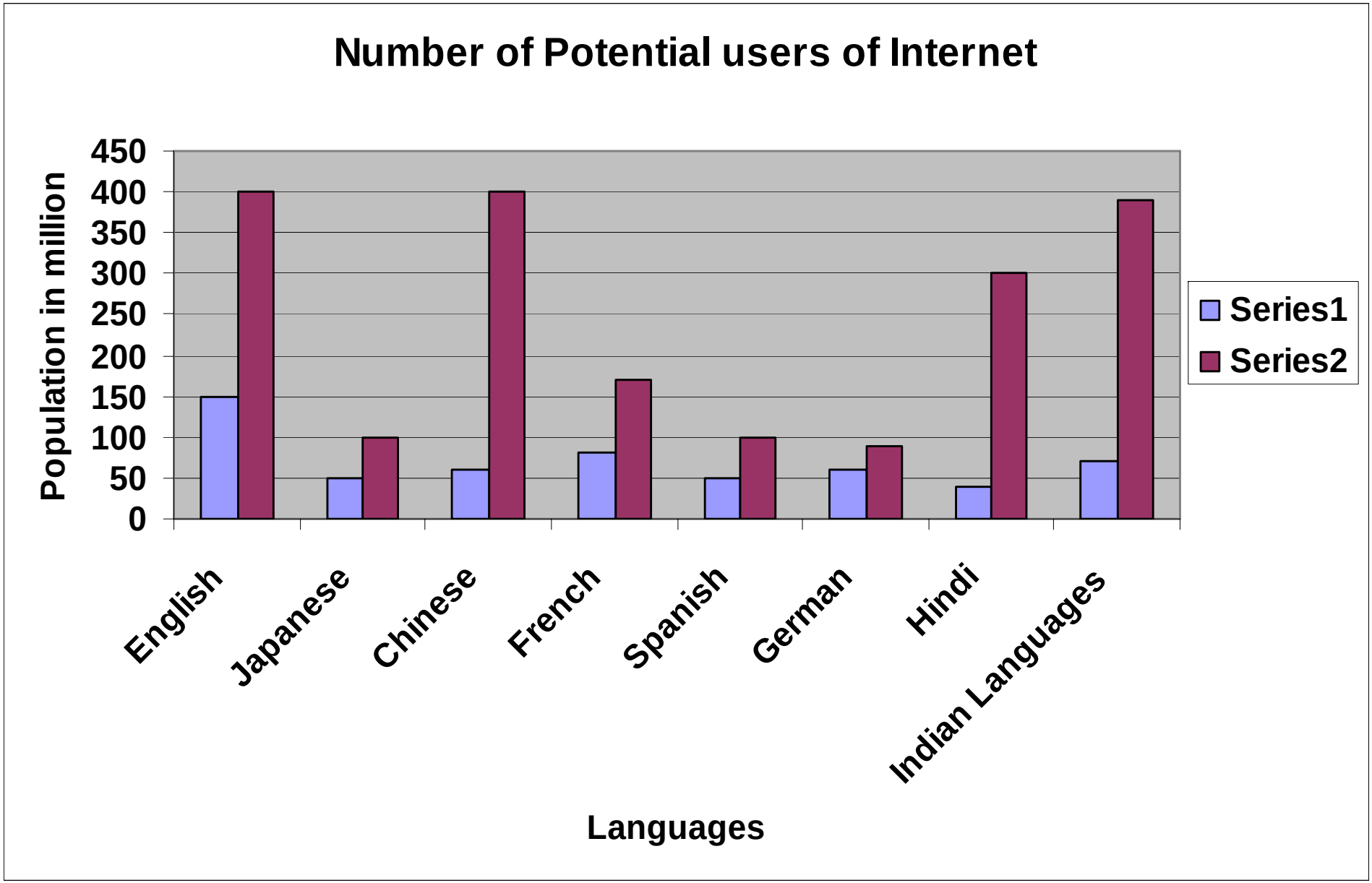
Wernicke's area: Region located posteriorly in the left hemisphere in the superior temporal gyrus. It is responsible for comprehension of spoken words and sentences. This area is named after Carl Wernicke (in 1874)

Roles of Broca's and Wernicke's areas

- Broadly, Broca's area is concerned with Grammar while Wernick's area is concerned with semantics
- Damage to former interferes with grammar, e.g. role confusion with voice change: "Ram was seen by Shyam" interpreted as *Ram is the seer*
- Damage to Wernick's area: finds it difficult to put a name to an entity (which is a tough categorization task)
- Evidence of difference between humans and apes in the complexity of language processing: Frontal lobe heavily used in humans ("The brain differentiates human and non-human grammars: Functional localization and structural connectivity" (Volume 103, Number 7, Pages 2458-2463, February 14, 2006)).

MT is needed: Internet Accessibility Pattern

User Type (script)	% of World Population	% access to the Internet
Latin	39	84
Kanzi (CJK)	22	13
Arabic	9	1.2
Brahmi and Indic	22	0.3



No of Internet Users in the year 2001

No of Internet Users in the year 2010 (Projected)

Living Languages

Continent	**No of languages**
Africa	2092
Americas	1002
Asia	2269
Europe	239
Pacific	1310
Total	*6912*

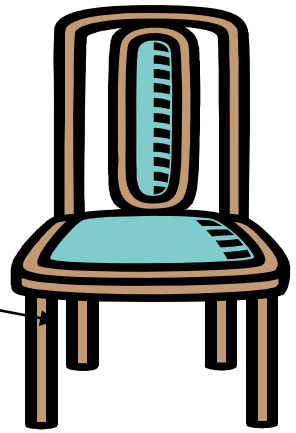
Stages and Challenges of NLP

NLP is concerned with *Grounding*

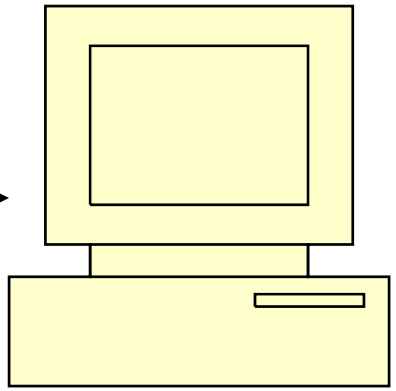
Ground the language into perceptual, motor and cognitive capacities.

Grounding

Chair



Computer

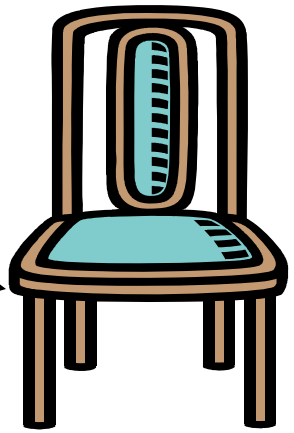


Grounding faces 3 challenges

- Ambiguity.
- Co-reference resolution (*anaphora* is a kind of it).
- Elipsis.

Ambiguity

Chair



Co-reference Resolution

Sequence of commands to the robot:

Place the wrench on the table.

Then paint it.

What does *it* refer to?

Elipsis

Sequence of command to the Robot:

Move the table to the corner.

Also the chair.

Second command needs completing by using the first part of the previous command.

Stages of processing (*traditional view*)

- Phonetics and phonology
- Morphology
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmatics
- Discourse

Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatraa* vs. *maatra (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come)* or *aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Phrase boundary
    - Milind Sohoni's mail announcing this seminar: *mtech1 students are especially exhorted to attend as such seminars are integral to one's post-graduate education*
  - Disfluency: *ah, um, ahem etc.*

Morphology

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys);* Gender marking (czar-czarina)
- Verbs: Tense (*stretch-stretched);* Aspect (*e.g. perfective sit-had sat*); Modality (e.g. *request khaanaa→ khaaiie)*
- First crucial first step in NLP
- Languages rich in morphology: e.g., Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing
- A task of interest to computer science: *Finite State Machines for Word Morphology*

Lexical Analysis

- Essentially refers to dictionary access and obtaining the properties of the word

  e.g. dog

  noun (lexical property)
  take-'s'-in-plural (morph property)
  animate (semantic property)
  4-legged (-do-)
  carnivore (-do)

Challenge: *Lexical or word sense disambiguation*

Lexical Disambiguation

First step: *part of Speech Disambiguation*

- *Dog* as a *noun* (animal)
- *Dog* as a verb (*to pursue)*

Sense Disambiguation

- *Dog (*as *animal)*
- *Dog (*as *a very detestable person)*

Needs word relationships in a context

- *The chair emphasised the need for adult education*

Very common in day to day communications and can occur in the form of single or multiword expressions

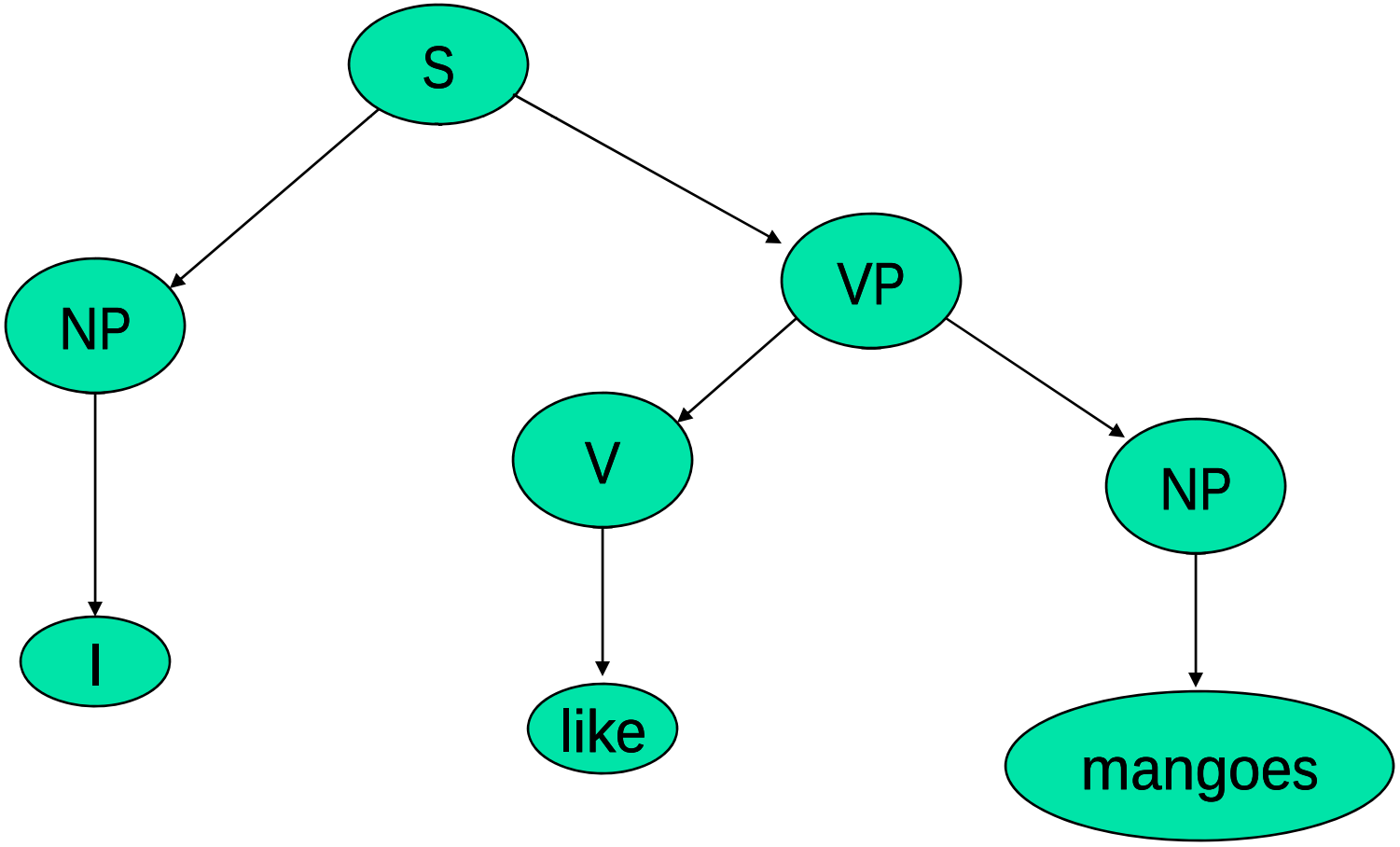
e.g., **Ground breaking ceremony** (Prof. Ranade's email to faculty 14/9/07)

Technological developments bring in new terms, additional meanings/nuances for existing terms

- Justify as in *justify the right margin* (word processing context)
- *Xeroxed:* a new verb
- *Digital Trace:* a new expression
- *Communifaking:* pretending to talk on mobile when you are actually not
- *Discomgooglation:* anxiety/discomfort at not being able to access internet
- *Helicopter Parenting*: over parenting

Syntax

Structure Detection



Parsing Strategy

- Driven by grammar
  - S-> NP VP
  - NP-> N | PRON
  - VP-> V NP | V PP
  - N-> Mangoes
  - PRON-> I
  - V-> like

Challenges: Structural Ambiguity

- Scope
  - *The old men and women were taken to safe locations*

  (old men and women) vs. *((old men) and women)*

  Seen in Amman airport: *No smoking areas will allow Hookas inside*
- Preposition Phrase Attachment
  - *I saw the boy with a telescope*

    (who has the *telescope?*)
  - I saw the mountain with a telescope

    (world knowledge: *mountain* cannot be an *instrument of seeing*)
  - I saw the boy with the pony-tail

    (world knowledge: *pony-tail* cannot be an *instrument of seeing*)

  Very ubiquitous: today's newspaper headline "*20 years later, BMC pays father 20 lakhs for causing son's death*"

Structural Ambiguity...

- Overheard
  - *I did not know my PDA had a phone for 3 months*
- An actual sentence in the newspaper
  - *The camera man shot the man with the gun when he was near Tendulkar*

Headache for parsing: Garden Path sentences

- Consider
  - *The horse raced past the garden* (sentence complete)
  - *The old man* (phrase complete)
  - *Twin Bomb Strike in Baghdad* (news paper heading: complete)

Headache for Parsing

- Garden Pathing
  - *The horse raced past the garden fell*
  - *The old man the boat*
  - *Twin Bomb Strike in Baghdad kill 25* (Times of India 5/9/07)

Semantic Analysis

- Representation in terms of
  - Predicate calculus/Semantic Nets/Frames/Conceptual Dependencies and Scripts
- *John gave a book to Mary*
  - Give action: Agent: John, Object: Book, Recipient: Mary
- Challenge: ambiguity in semantic role labeling
  - *(Eng) Visiting aunts can be a nuisance*
  - *(Hin) aapko mujhe mithaai khilaanii padegii (ambiguous in Marathi and Bengali too; not in Dravidian languages)*

Pragmatics

- Very hard problem
- Model user intention
  - *Tourist (in a hurry, checking out of the hotel, motioning to the service boy): Boy, go upstairs and see if my sandals are under the divan. Do not be late. I just have 15 minutes to catch the train.*
  - *Boy (running upstairs and coming back panting): yes sir, they are there.*
- World knowledge
  - *WHY INDIA NEEDS A SECOND OCTOBER (*ToI, 2/10/07, yesterday)

Discourse

Processing of *sequence* of sentences

Mother to *John*:

John go to school. It is open today. Should you bunk? Father will be very angry.

Ambiguity of *open*

bunk what?

Why will the father be angry?

Complex chain of reasoning and application of world knowledge

(*father will not be angry if somebody else's son bunks the school)*

Ambiguity of *father*

father as *parent*

or

father as *headmaster*

Complexity of Connected Text

John was returning from school dejected – today was the math test

He couldn't control the class

Teacher shouldn't have made him responsible

After all he is just a janitor

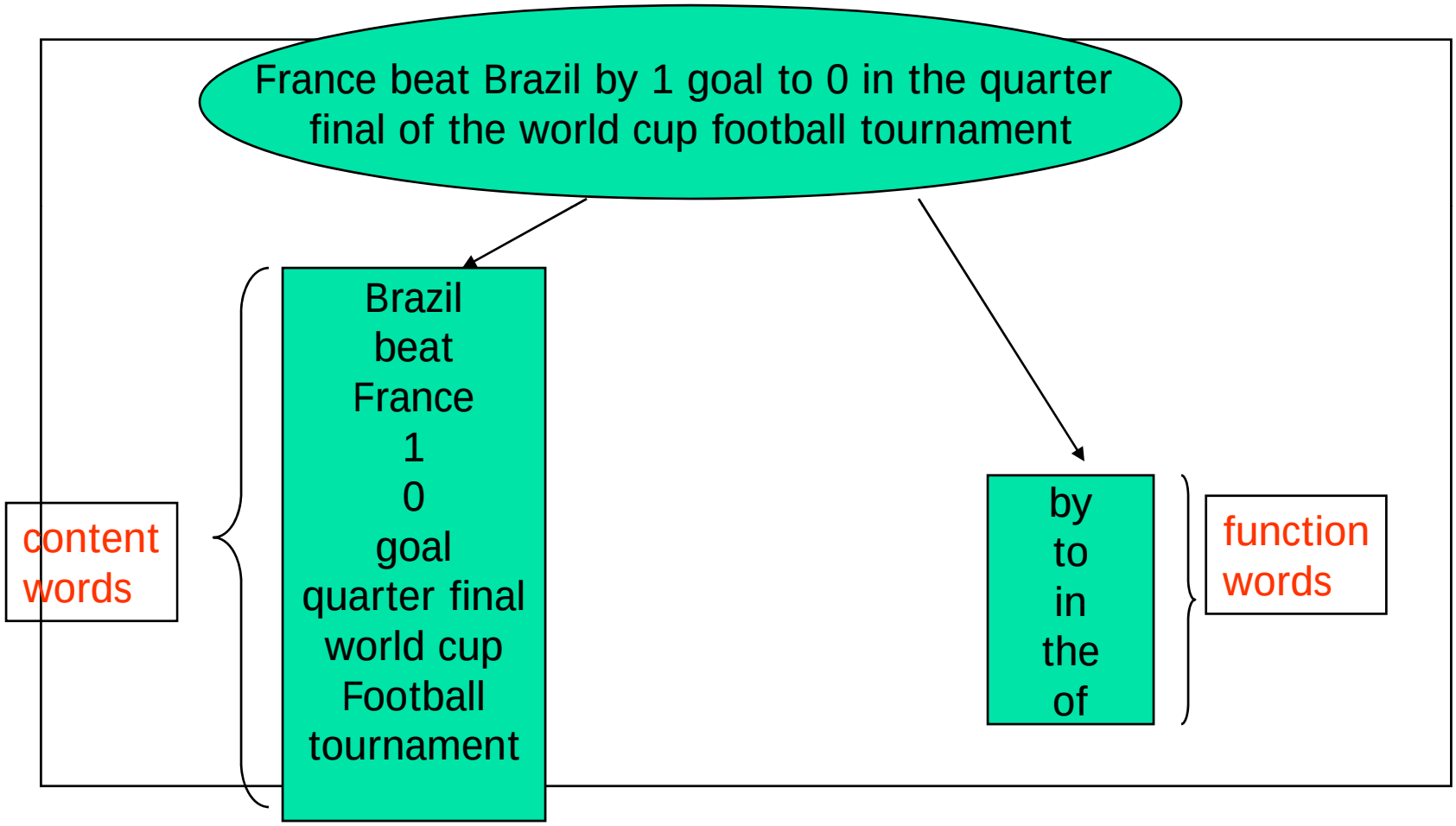
ML-NLP

NLP as an ML task

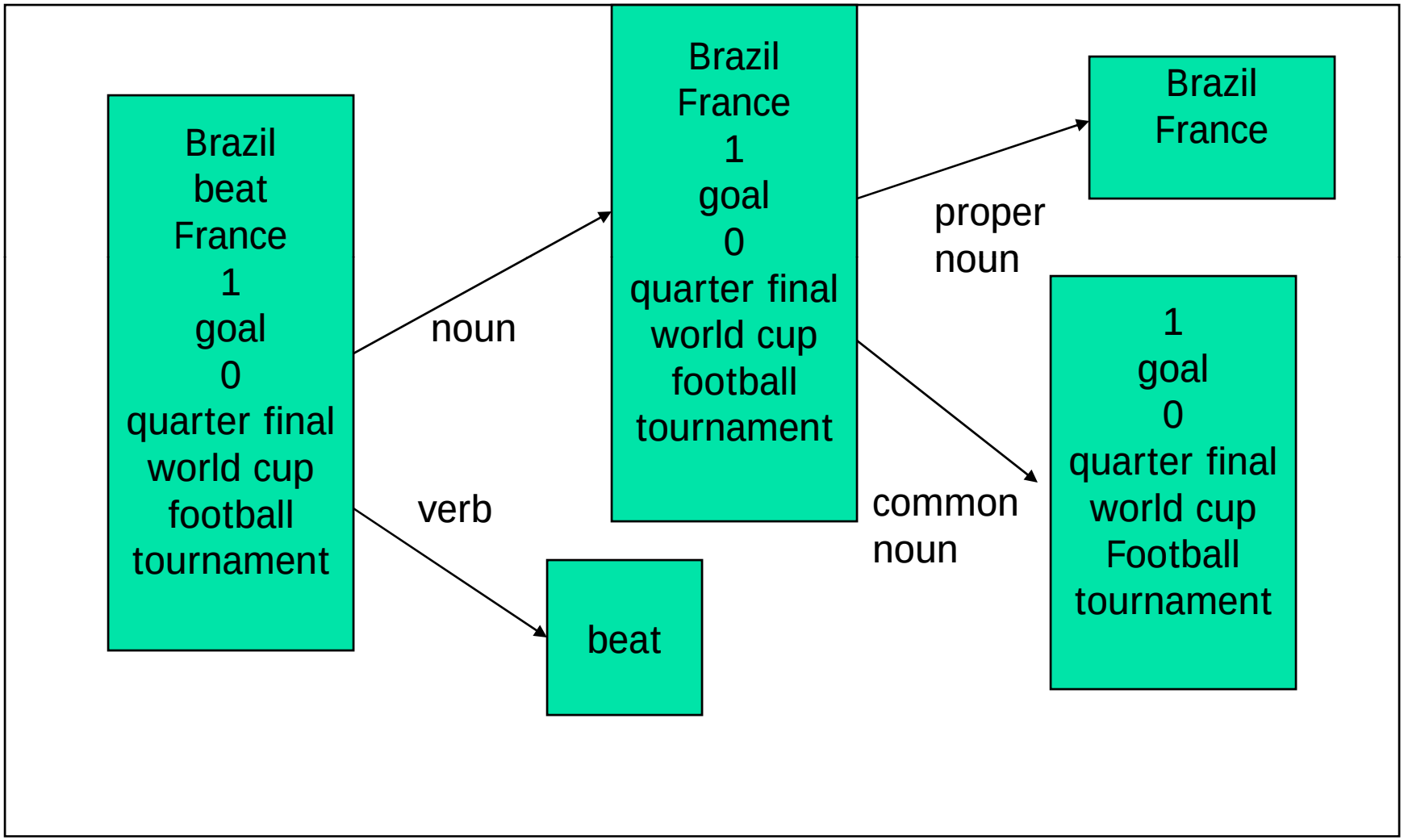
- France *beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament.* *(English)*

- *braazil ne phraans ko vishwa kap phutbal spardhaa ke kwaartaar phaainal me 1-0 gol ke baraabarii se haraayaa. (Hindi)*

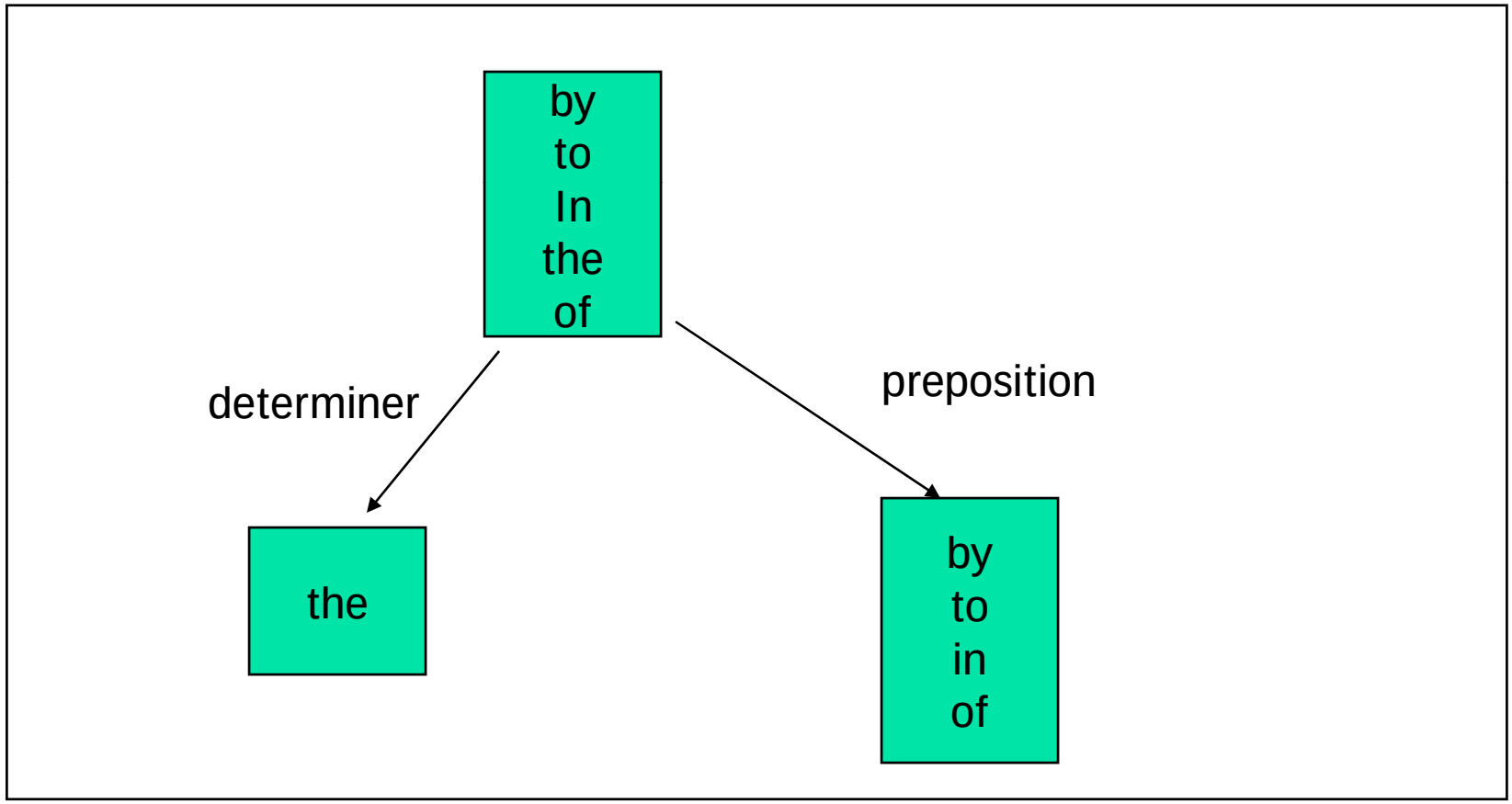
Categories of the Words in the Sentence



Further Classification *1/2*



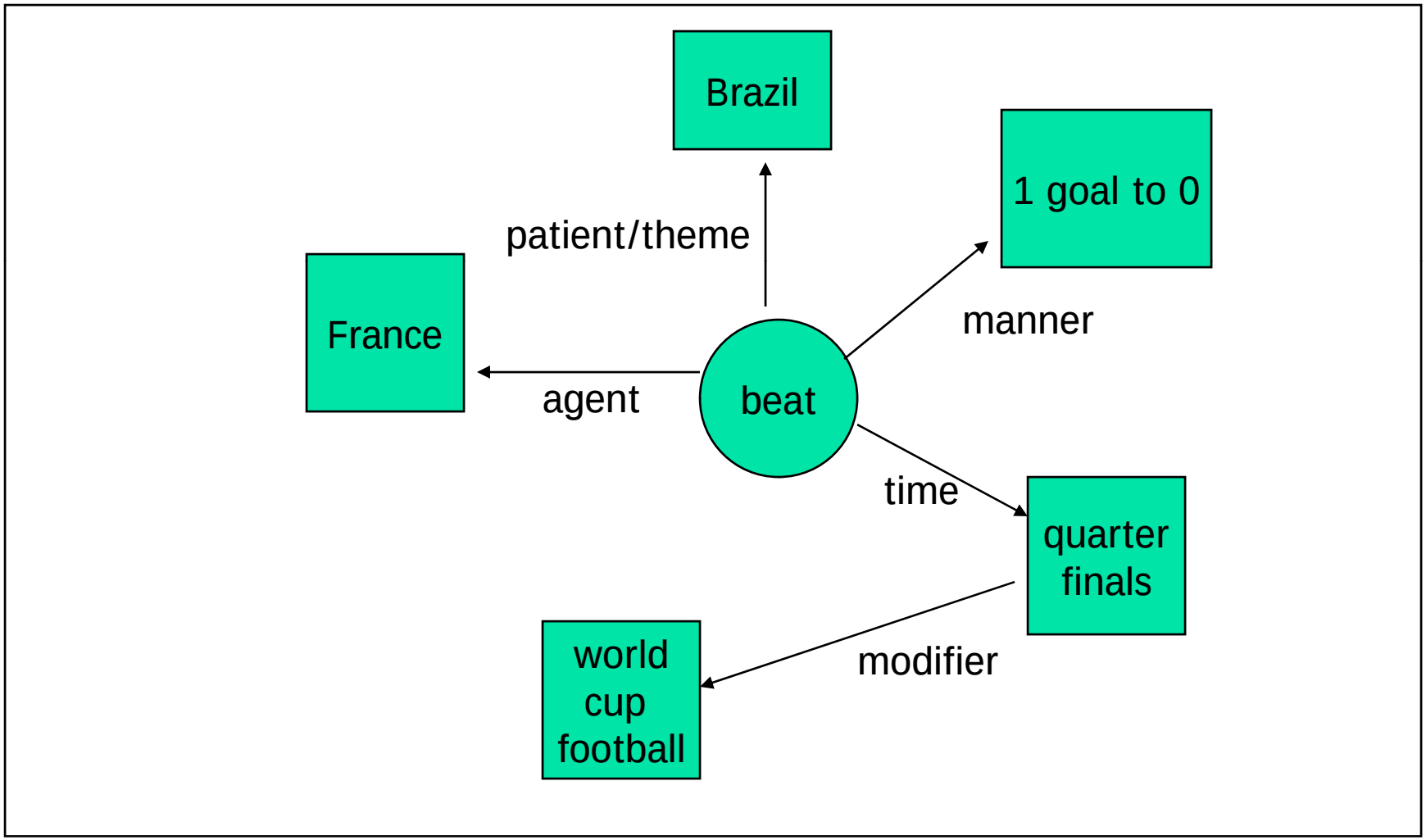
Further Classification *2/2*



Why all this?

- Fundamental and ubiquitous information need
  - *who did what*
  - *to whom*
  - *by what*
  - *when*
  - *where*
  - *in what manner*

Semantic roles



Semantic Role Labeling: *a classification task*

- *France beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament*
  - Brazil: *agent or object?*
  - Agent: *Brazil or France or Quarter Final or World Cup?*
- *Given an entity, what role does it play?*
- *Given a role, it is played by which entity?*

A lower level of classification: Part of Speech (POS) Tag Labeling

- *France beat Brazil by 1 goal to 0 in the quarter-final of the world cup football tournament*
  - beat: *verb of noun (heart beat, e.g.)?*
  - Final: *noun or adjective?*

Uncertainty in classification: **Ambiguity**

- *Visiting aunts can be a nuisance*
  - Visiting:
    - *adjective or gerund* (POS tag ambiguity)
  - Role of *aunt:*
    - *agent of visit* (aunts are visitors)
    - *object of visit* (aunts are being visited)
- Minimize uncertainty of classification with **cues** from the sentence

What *cues?*

- Position with respect to the verb:
  - *France* to the left of *beat* and *Brazil* to the right: agent-object role marking (English)
- Case marking:
  - *France ne (Hindi); ne (Marathi): agent role*
  - *Brazil ko (Hindi); laa (Marathi): object role*
- Morphology: *haraayaa (hindi); haravlaa (Marathi):*
  - *verb POS tag as indicated by the distinctive suffixes*

Cues are like *attribute-value pairs* prompting machine learning from NL data

- Constituent ML tasks
  - Goal: classification or clustering
  - Features/attributes (word position, morphology, word label *etc.*)
  - Values of features
  - Training data (corpus: annotated or un-annotated)
  - Test data (test corpus)
  - Accuracy of decision (precision, recall, F-value, MAP *etc.*)
  - Test of significance (sample space to generality)

What is the output of an ML-NLP System *(1/2)*

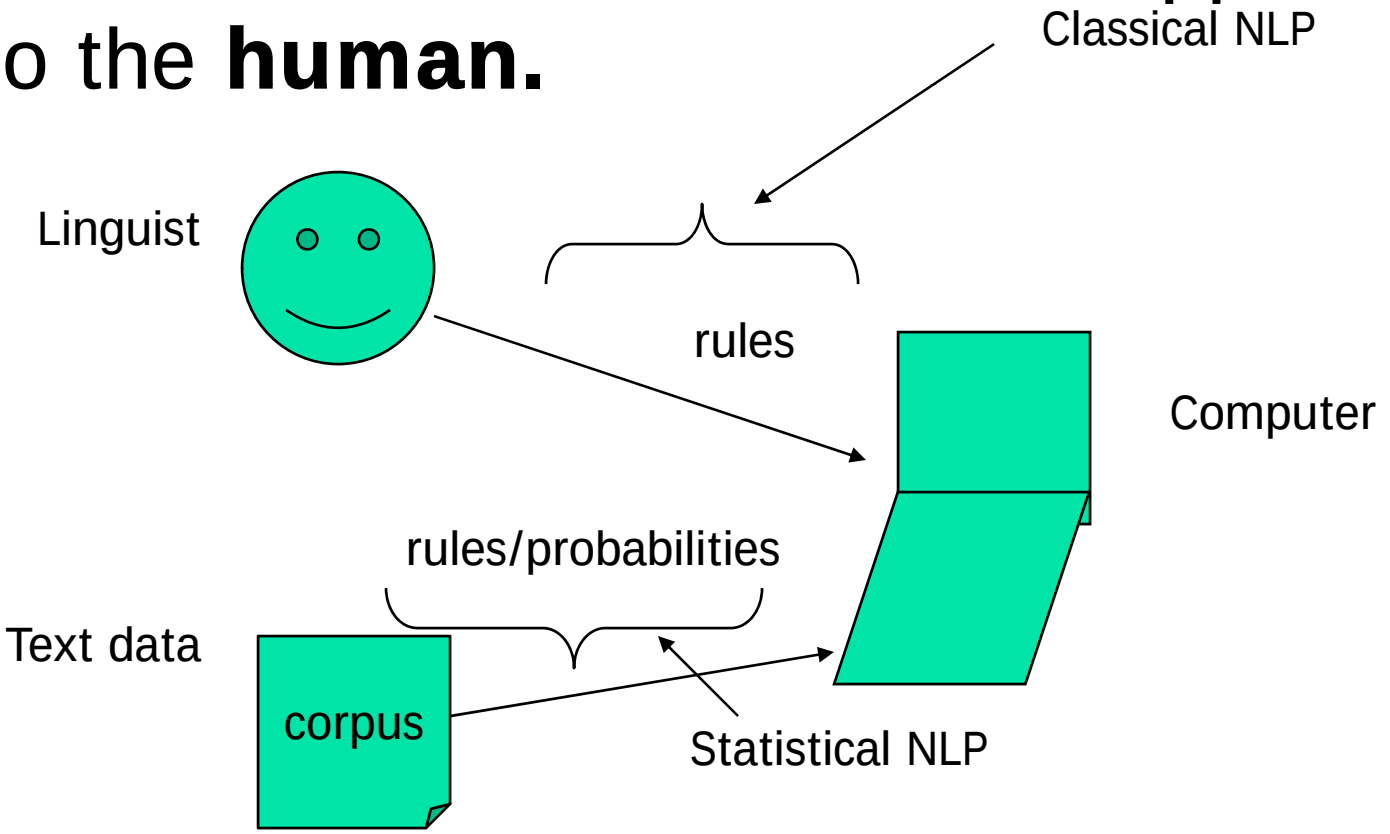
- Option 1: A set of rules, *e.g.*,
  - *If the word to the left of the verb is a noun and has animacy feature, then it is the likely **agent** of the action denoted by the verb.*
    - *The child broke the toy* (*child* is the agent)
    - *The window broke* (*window* is not the agent; inanimate)

What is the output of an ML-NLP System *(2/2)*

- Option 2: a set of probability values
  - *P(agent|word is to the left of verb and has animacy) > P(object|word is to the left of verb and has animacy)> P(instrument|word is to the left of verb and has animacy) etc.*

How is this different from classical NLP

- The burden is on the **data** as opposed to the **human.**



Classification appears as sequence labeling

A set of Sequence Labeling Tasks: *smaller to larger units*

- *Words*:
  - Part of Speech tagging
  - Named Entity tagging
  - Sense marking
- *Phrases*: Chunking
- *Sentences*: Parsing
- *Paragraphs*: Co-reference annotating

Example of word labeling: POS Tagging

<s>
Come September, and the IIT campus is abuzz with new and returning students.
</s>

<s>
Come_VB September_NNP ,_, and_CC the_DT IIT_NNP campus_NN is_VBZ abuzz_JJ with_IN new_JJ and_CC returning_VBG students_NNS ._.
</s>

Example of word labeling: Named Entity Tagging

<month_name>
September
</month_name>

<org_name>
IIT
</org_name>

Example of word labeling: Sense Marking

Word	Synset	WN-synset-no
come	*{arrive, get, come}*	*01947900*
	.	
	.	
	.	
abuzz	*{abuzz, buzzing, droning}*	*01859419*

Example of phrase labeling: Chunking

Come July, and the IIT campus is

abuzz with new and returning students .

Example of Sentence labeling: Parsing

[$_{S1}$[$_{S}$[$_{S}$[$_{VP}$[$_{VB}$Come][$_{NP}$[$_{NNP}$July]]]]

[$_{,}$,]

[$_{CC}$ and]

[$_{S}$ [$_{NP}$ [$_{DT}$ the] [$_{JJ}$ UJF] [$_{NN}$ campus]]

[$_{VP}$ [$_{AUX}$ is]

[$_{ADJP}$ [$_{JJ}$ abuzz]

[$_{PP}$[$_{IN}$ with]

[$_{NP}$[$_{ADJP}$ [$_{JJ}$ new] [$_{CC}$ and] [$_{VBG}$ returning]]

[$_{NNS}$ students]]]]]]

[$_{.}$.]]]