Indian Language Lexical Resources

Pushpak Bhattacharyya

Computer Science and Engg. Dept.

Indian Institute of Technology Bombay

The Genesis

Universal Networking Language

Common language for computers to express information written in natural language

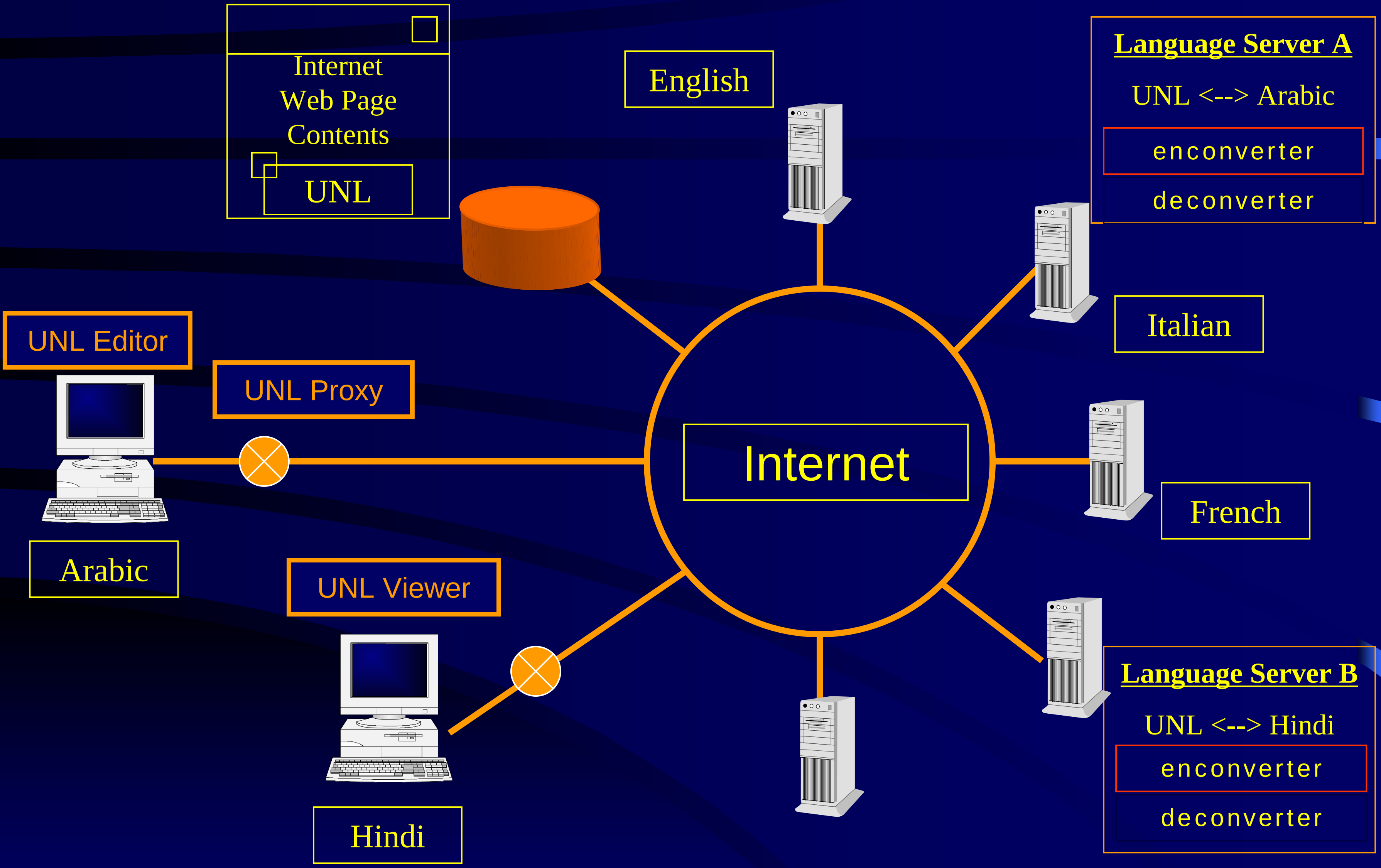
•Large International Effort, IIT Bombay became a part in 1996

•Application:

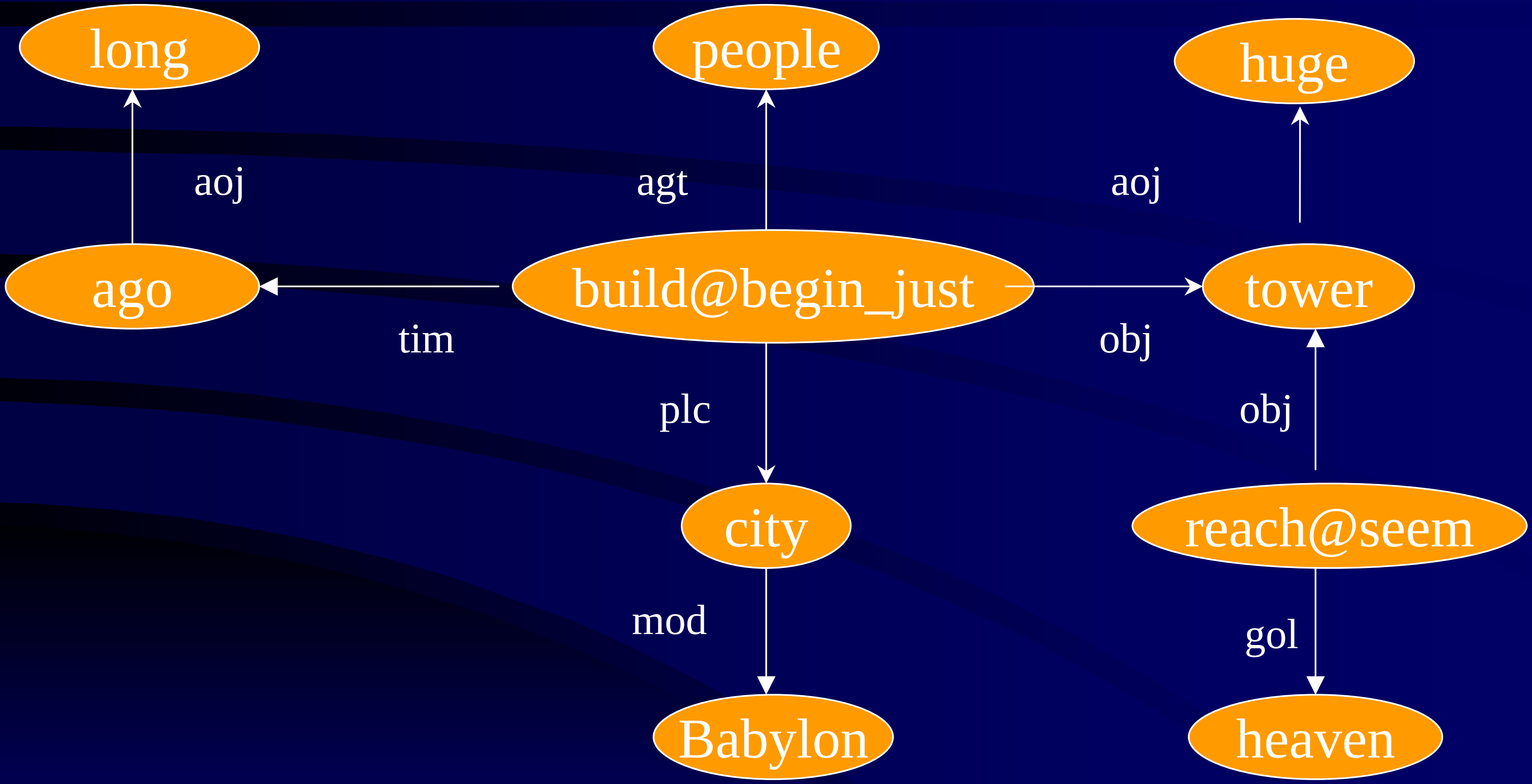
Electronic language to overcome language barrier

Information Distribution System

The UNL System



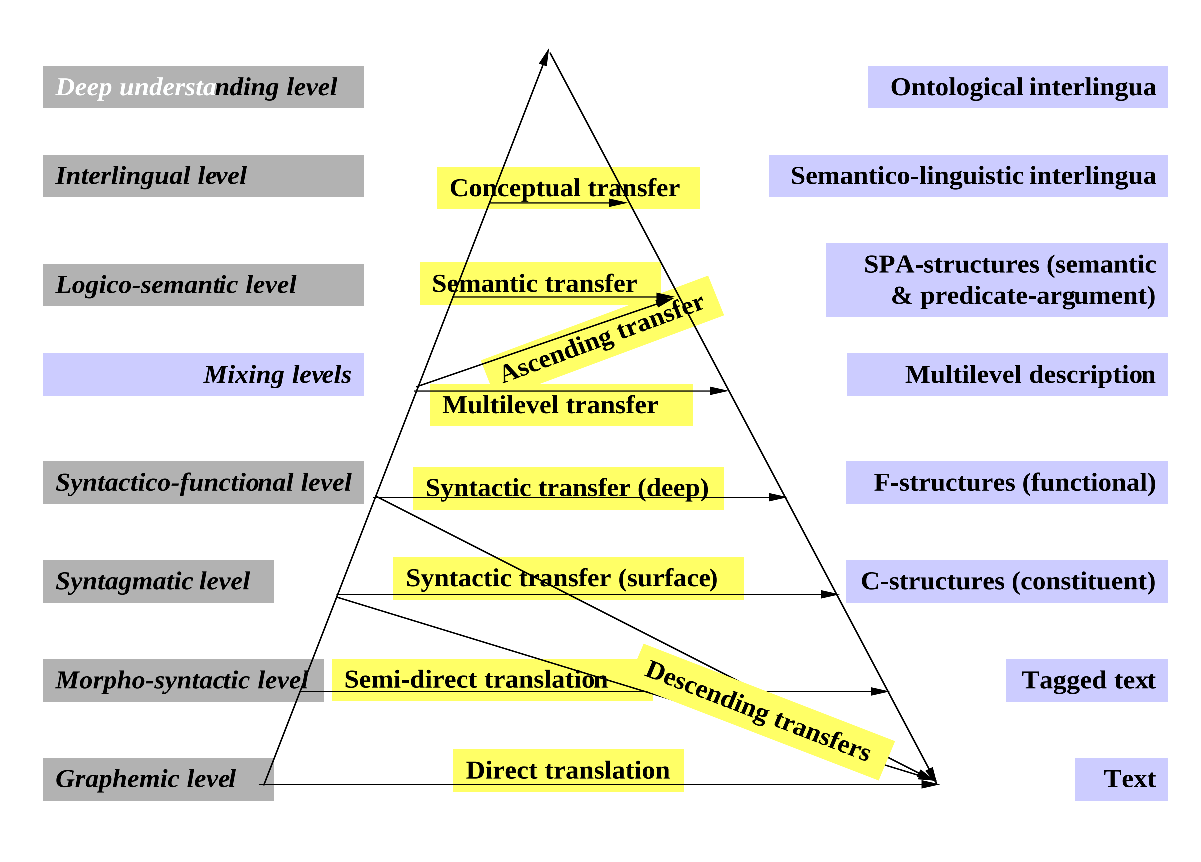
Long ago, in the city of Babylon, the people began to build a huge tower, which seemed about to reach the heavens.



Kinds of MT Systems

(point of entry from source to the target text)

fwd



Deep understanding level

Ontological interlingua

Interlingual level

Conceptual transfer

Semantico-linguistic interlingua

Logico-semantic level

Semantic transfer

SPA-structures (semantic & predicate-argument)

Ascending transfer

Mixing levels

Multilevel transfer

Multilevel description

Syntactico-functional level

Syntactic transfer (deep)

F-structures (functional)

Syntagmatic level

Syntactic transfer (surface)

C-structures (constituent)

Morpho-syntactic level

Semi-direct translation

Descending transfers

Tagged text

Graphemic level

Direct translation

Text

Universal Word

- Vocabulary of the UNL
- A UW represents a concept

  1) Basic UW

  2) Restricted UW

ex) spring(icl>tool)

spring(icl>season)

Languages Covered

- 6 UN official Languages.

  Arabic, Chinese, English, French, Spanish, Russian

- Other languages

  German, Greek, Hindi, Indonesian, Italian, Japanese, Korean, Mongol, Latvia, Portuguese, Thai

Lexicon

- Format of the Lexicon

  [HW] {ID} “UW” (ATTRIB1, ATTRIB2,….) <FLG,FRE,PRI>;

- Examples:

  [Bird] {} “bird (icl>animal>concrete volitional thing)” (N,ANI,SG,CONCRETE)<E,0,0>;

  [try] {} “try(icl>event)” (V,PRES,SIMPL,DISC)<E,0,0>;

  [from] {} “from” (PRE,INI,PLC,TIME,INEFF)<E,0,0>;

  [What] {} “what”(PRON,INTER,WH-WORD)<E,0,0>;

From the English wordnet: Bird

- The noun bird has 5 senses (first 2 from tagged texts)
-
- 1. (31) bird -- (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
- 2. (1) bird, fowl -- (the flesh of a bird or fowl (wild or domestic) used as food)
- 3. dame, doll, wench, skirt, chick, bird -- (informal terms for a (young) woman)
- 4. boo, hoot, Bronx cheer, hiss, raspberry, razzing, razz, snort, bird -- (a cry or noise made to express displeasure or contempt)
- 5. shuttlecock, bird, birdie, shuttle -- (badminton equipment consisting of a ball of cork or rubber with a crown of feathers)

- The verb bird has 1 sense (no senses from tagged texts)
-
- 1. bird, birdwatch -- (watch and study birds in their natural habitat)

Role of Semantic Knowledge

Mary eats noodles with a fork.	ins(eat(icl>do), fork(icl>thing))
Mary eats noodles with John.	cag(eat(icl>do),John)
Ram eats noodles with vegetables.	cob(eat(icl>do),vegetable.@pl)
A demon eats noodles with a goat.	cob?? Or cag??

Semantic Attributes are used to solve such PP-Attachment ambiguities and decide UNL relations.

Example:

The words in above sentences are classified as follows:

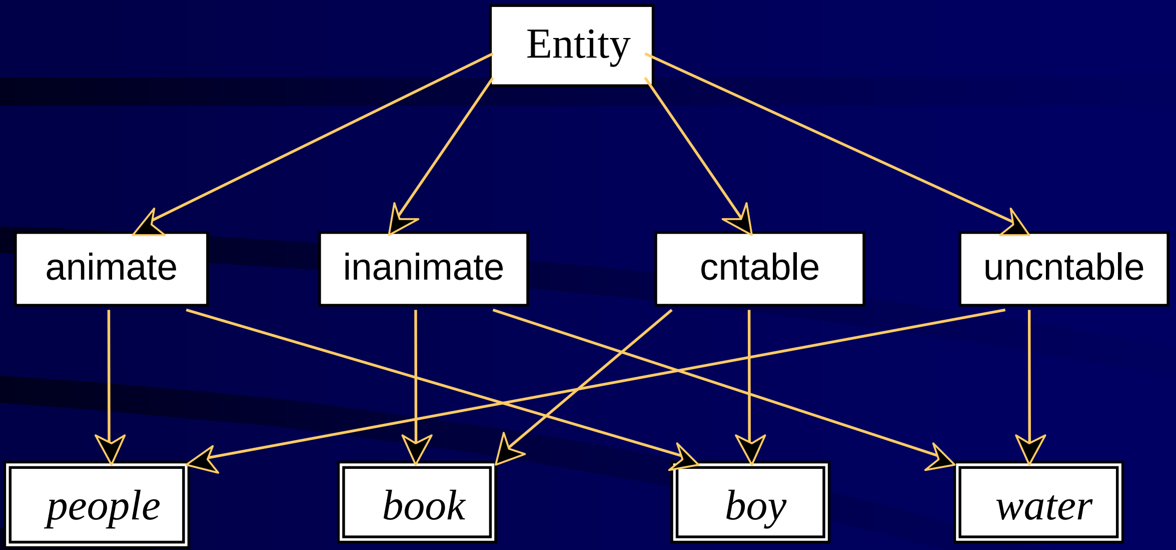
Fork is classified under inanimate things that can be used as instruments.

Mary and John are classified as animate things that are humans and Omnivores.

Demon is similarly classified as animate things that are Omnivores.

Noodles and vegetables are classified as inanimate things that are edible.

Ontology for enriching Lexicon



- Add attributes to the lexicon by traversing the DAG

[people]{}”people(icl>community)”(N,ANI,UNCNT)<E,0,0>
[book]{}”book(icl>thing)”(N,INANI,CNT)<E,0,0>
[boy]{}”boy(icl>human)”(N,ANI,CNT)<E,0,0>
[water]{}”water(icl>fluid)”(N,INANI,UNCNT)<E,0,0>

Current potential of the System

The soldier went away to the totally deserted desert to desert the house in the desert

```
;======================= UNL ======================
;The soldier went away to the totally deserted desert to desert the house in the desert
[S]
mod(deserted(icl>vacant):11,total(icl>complete):0T)
aoj(deserted(icl>vacant):11,desert(icl>landscape):1A.@def)
plc(go(icl>event):0C.@entry.@past.@pred,away(icl>logical place):0H)
obj(desert(icl>do):1K.@present.@pred,house(icl>place):1V.@def)
plc(desert(icl>do):1K.@present.@pred,desert(icl>landscape):28.@def)
plt(go(icl>event):0C.@entry.@past.@pred,desert(icl>landscape):1A.@def)
pur(go(icl>event):0C.@entry.@past.@pred,desert(icl>do):1K.@present.@pred)
agt(go(icl>event):0C.@entry.@past.@pred,soldier(icl>human):04.@def)
[/S]
;==================================================
```

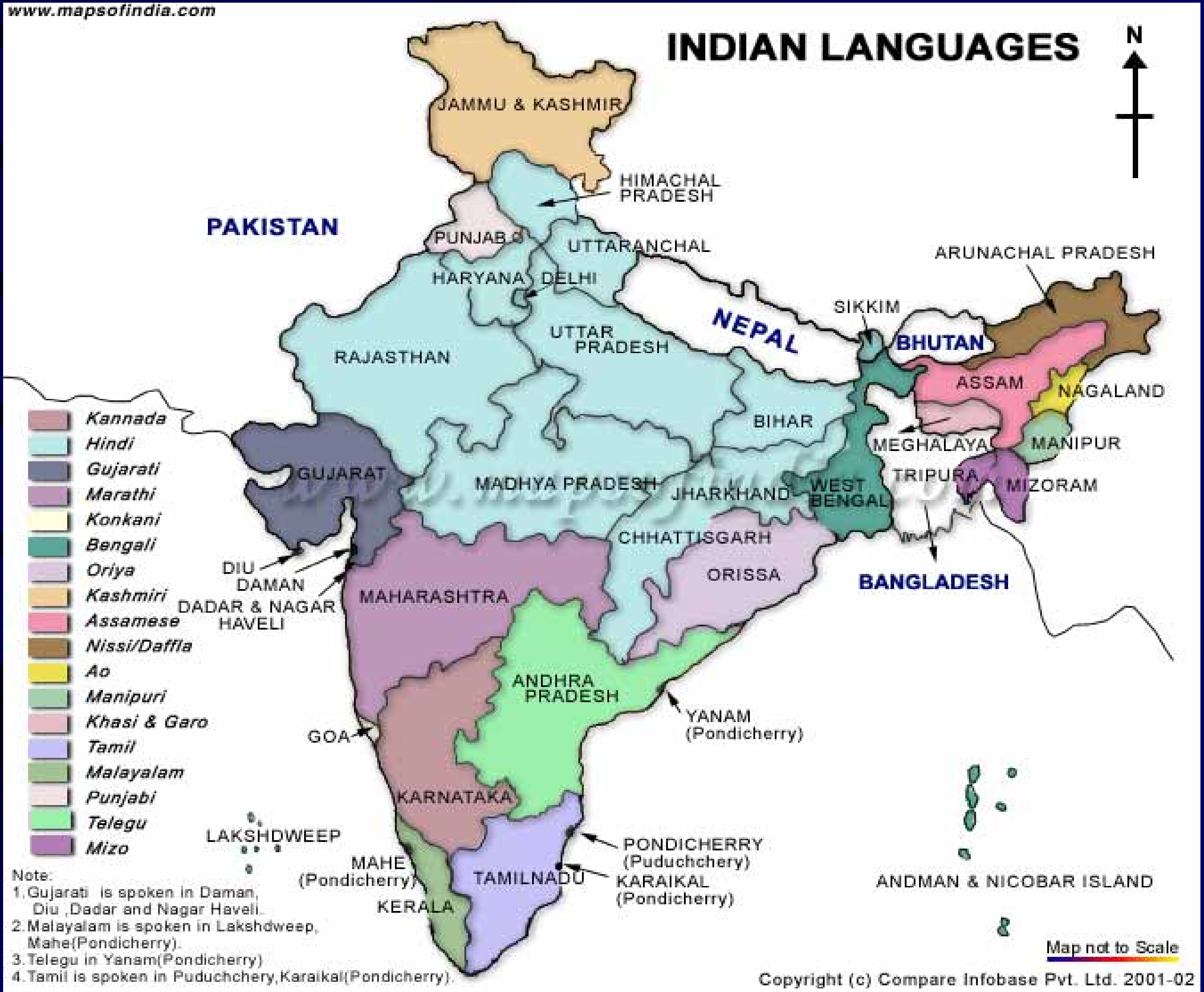
Hindi Wordnet at IIT Bombay

Current Team:

Pushpak Bhattacharyya, Prabhakar Pande, Laxmi Kashyap, Devendra Kairwan, Salil Joshi, Arun Karthikeyan, Prachur Goel

and many previous PhD, Masters and Bachelor Students and Research Staff

Great Language Diversity of India



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).

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

Major Language Processing Initiatives

- Mostly from the Government: *Ministry of IT, Ministry of Human Resource Development, Department of Science and Technology*
- Recently great drive from the industry: NLP efforts with Indian language in focus
  - *Google, Microsoft, IBM Research Lab*
  - *Yahoo, TCS*

IIT Bombay Natural Language Processing Group heavily supported by Government and Industry

What is Hindi Wordnet

- Wordnet – A lexical database
- Hindi Wordnet Inspired by the English WordNet
- Built conceptually
- *Synsets* or the Synonymy Sets are the basic building blocks
- Different organizing principles for different syntactic categories

Example Entry in Hindi Wordnet

- Synset

  { , , , }

  {gaaya ,gauu, gaiyaa, dhenu}, Cow
- Gloss
  - Text definition

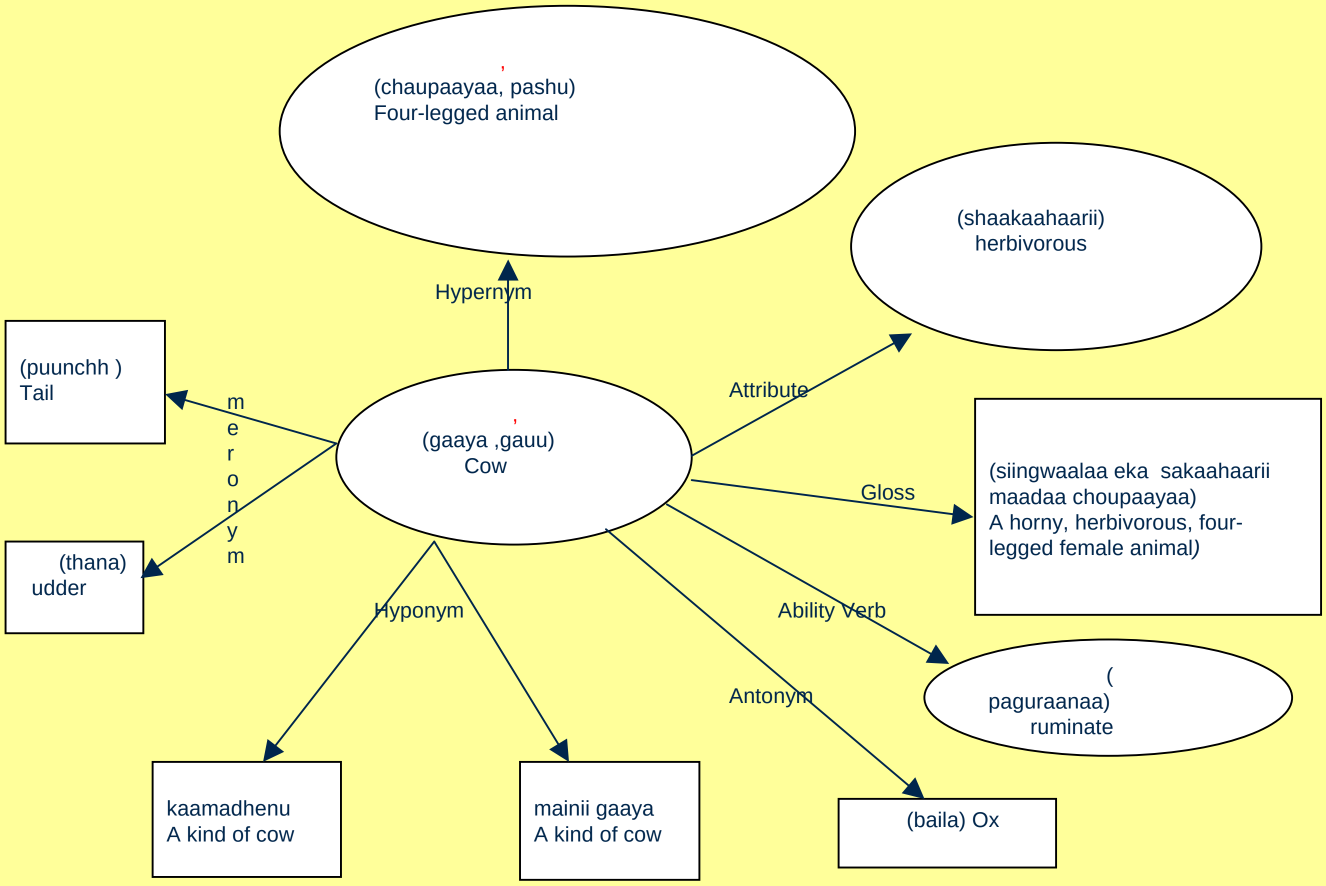
    (siingwaalaa eka shaakaahaarii maadaa choupaayaa)
    *(*a horny, herbivorous, four-legged female animal*)*
  - Example sentence

    (hinduu loga gaaya ko go maataa kahate hain evam usakii puujaa karate hain)
    (The Hindus considers cow as mother and worship it.)

Relations in Wordnet

- Synonymy
- Hypernymy / Hyponymy
- Antonymy
- Meronymy / Holonymy
- Gradation
- Entailment
- Troponymy

WordNet Sub-Graph: Hindi



Statistics

Synsets	33500
Unique Words	80400
Related Synsets	33500
Hindi-English Linked Synsets	13000
Hits	260000

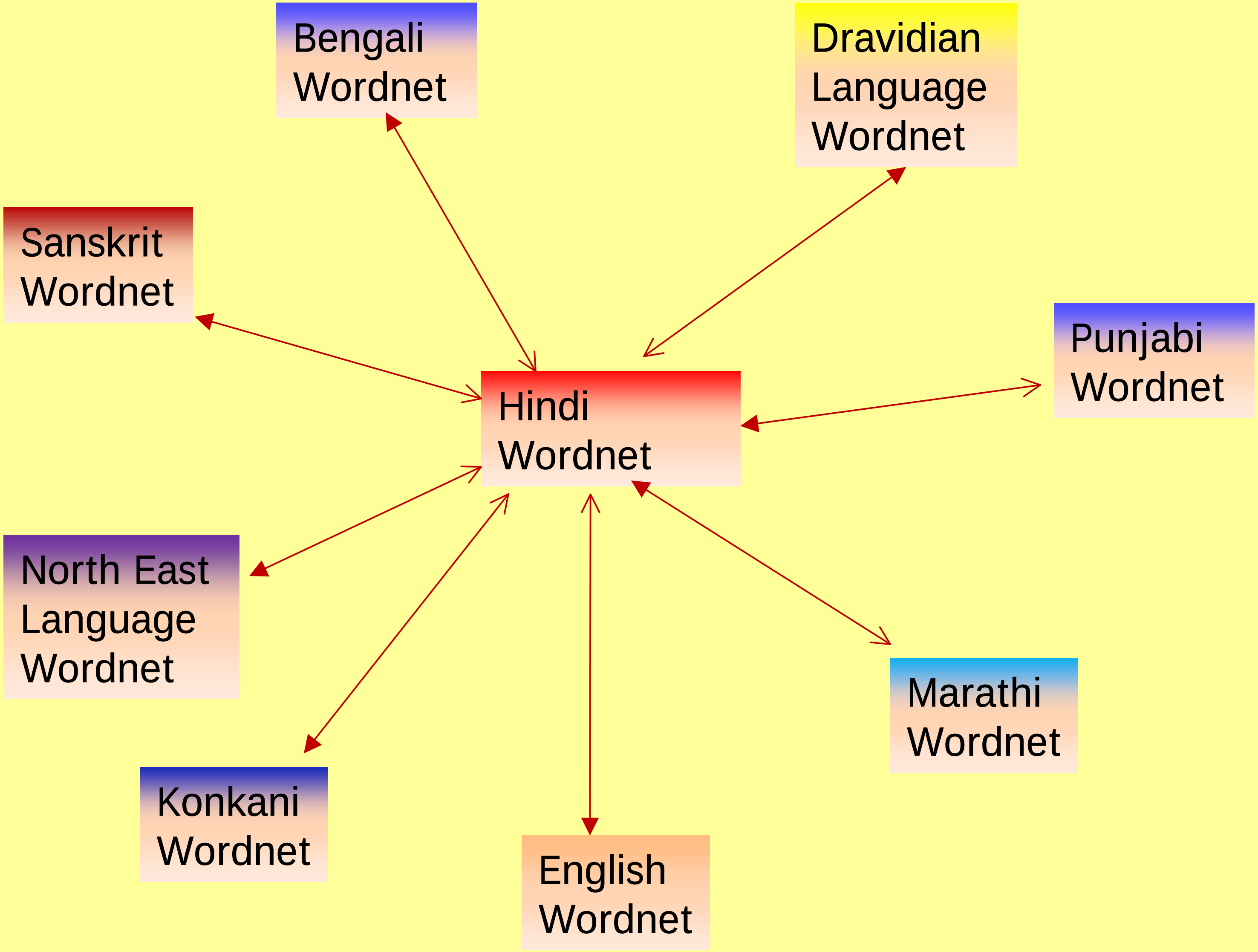
Impact, Use and Visibility of Hindi Wordnet

- Free download with API under *GPL*
- Available from LDC (linguistics data consortium), Upenn: topmost linguistic data repository in the worlds
- Commercial license purchased by Google for work on Indian language search engine
- Discussions on with Microsoft, Inforsys
- To be available from ELRA: language data repository of Europe
- Available from LDC-IL: LDC of India

Impact, Use and Visibility of created resources *(continued)*

- Daily reference form all over the world
- More than 2 Lakh hits so far since 2006
- More than 3000 downloads
- Pivot for wordnets of many Indian languages
- Base resource used by many researchers for IL work on translation, summarization, cross lingual search

Hindi Wordnet giving rise to other Indian Language wordnets



Bengali Wordnet

Dravidian Language Wordnet

Sanskrit Wordnet

Punjabi Wordnet

Hindi Wordnet

North East Language Wordnet

Marathi Wordnet

Konkani Wordnet

English Wordnet

Linked wordnets

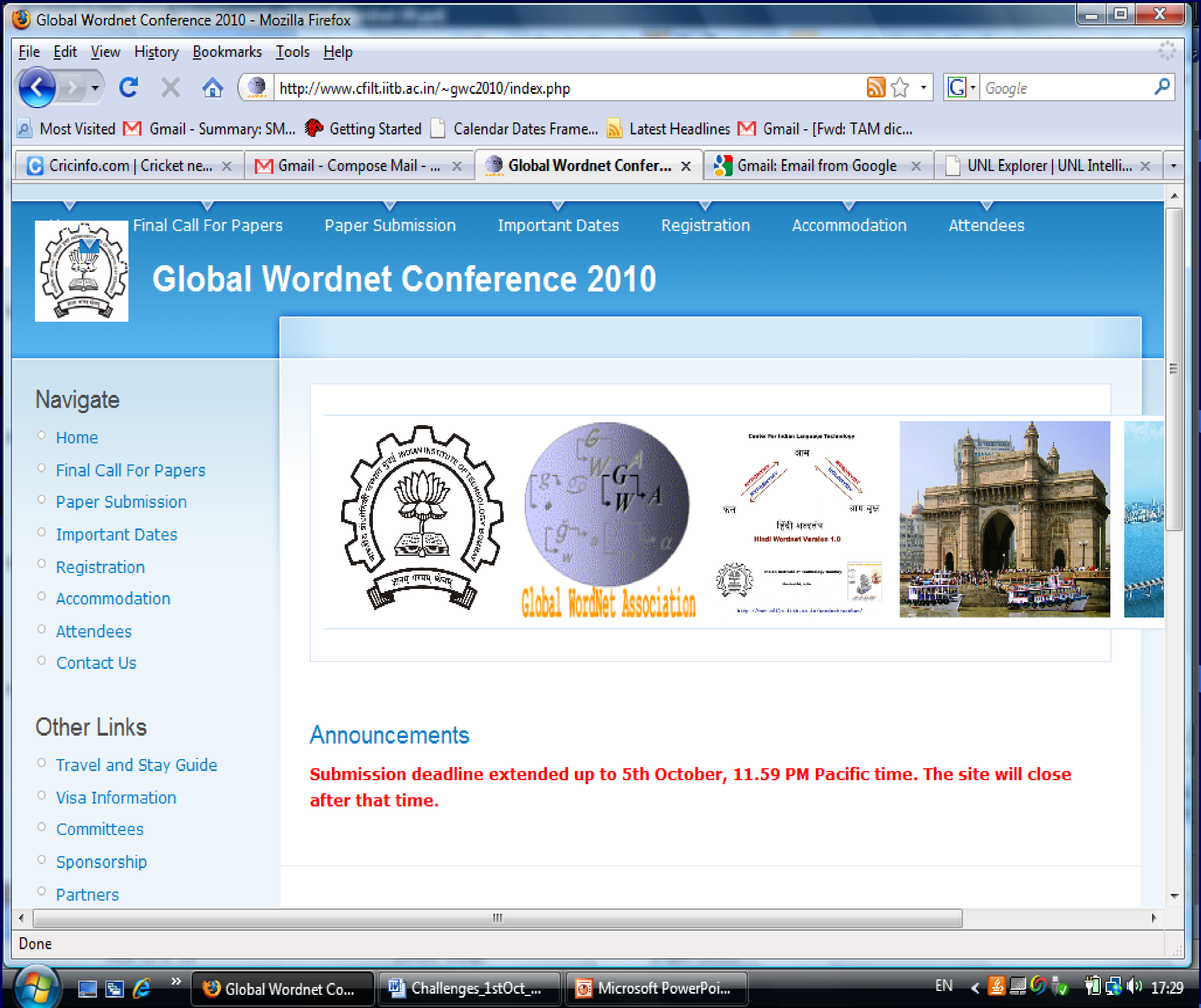
- Immense Lexical Resource
- Great benefits to machine translation, cross lingual search
- Very useful for language teaching, pedagogy, comparative linguistics
- Akin to Eurowordnet, but critical differences due to typical Indian language characteristics

Pan-India Dictionary Standard based on wordnet

Senses	Hindi	Marathi	Bangali	Oriya	Tamil
(W_1, W_2, W_3, W_4, W_5, W_6)	(W_1, W_2, W_3, W_4, W_5, W_6)	(W_1, W_2, W_3)	(W_1, W_2 , W_3)	(W_1, W_2, W_3, W_4)	(W_1, W_2, W_3)
(sun)	(सूर्‌य, सूरज, भानु, भास्‌कर, प्‌रभाकर, दनिकर, अंशुमान, अंशुमाली)	(सूर्‌य, भानु, दविाकर, भास्‌कर, रवि, दनिश, दनिमणी)	…	…	…
(cub, lad, laddie, sonny, sonny boy)	(, , , , , ,)	(, , ,)	. . .	. . .	. . .
(son, boy)	(, , , , , , , , ,)	(, , , ,)	. . .	. . .	. . .

International Global Wordnet Conference, Jan 31-Feb 4, 10

A major International Event Granted to IIT Bombay Because of The success Of Hindi Wordnet



Global Wordnet Conference 2010

Navigate

- Home
- Final Call For Papers
- Paper Submission
- Important Dates
- Registration
- Accommodation
- Attendees
- Contact Us

Other Links

- Travel and Stay Guide
- Visa Information
- Committees
- Sponsorship
- Partners

Announcements

Submission deadline extended up to 5th October, 11.59 PM Pacific time. The site will close after that time.

Conclusion

- It important to represent words accurately and with rich information

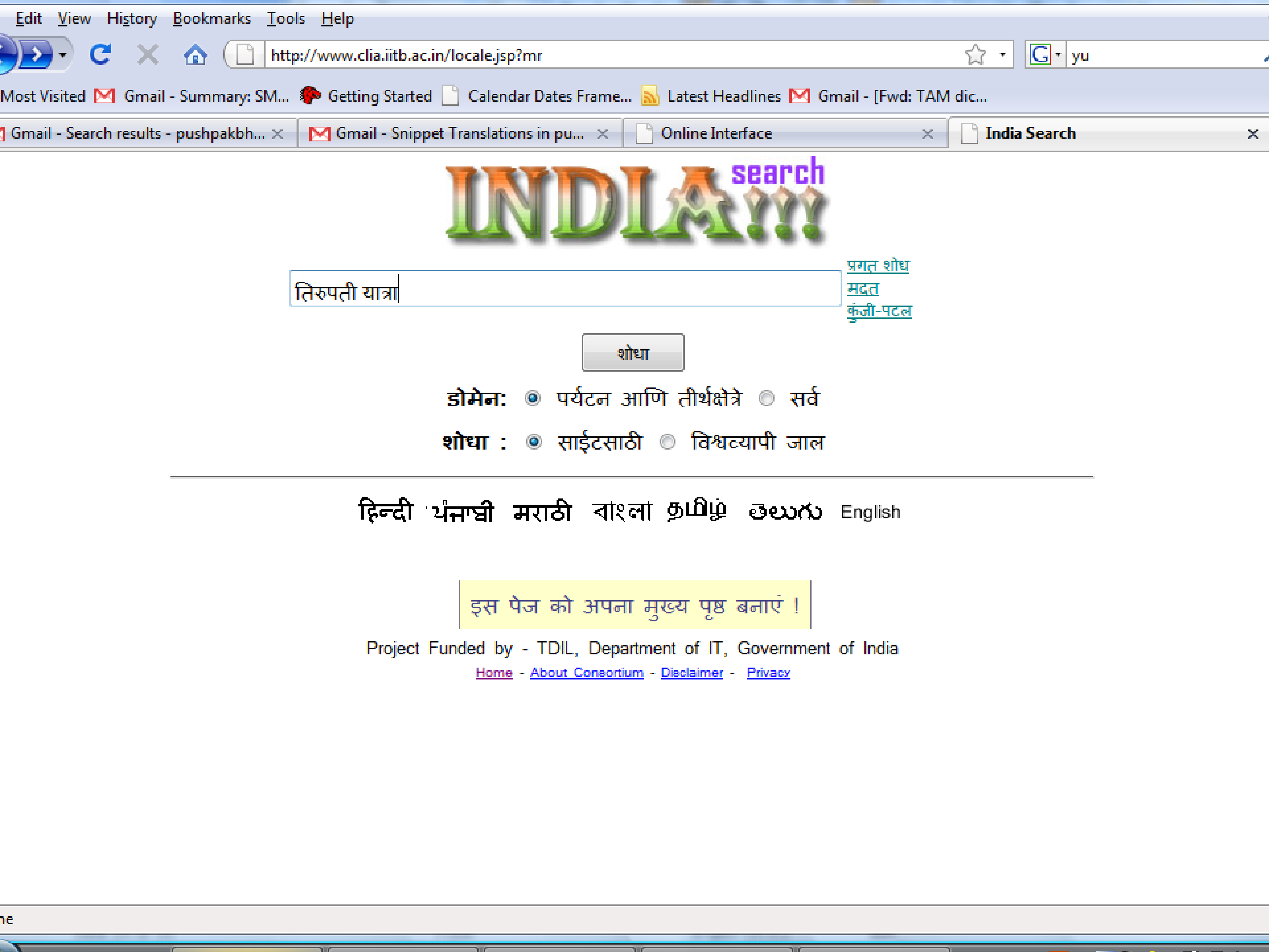
- Multifarious fundamental applications:

1. Information Extraction
2. Machine Translation on the internet
3. Text Mining
4. Cross Lingual Search

(PTO.)

Cross Lingual Search Portal

- www.clia.iitb.ac.in
- तरिुपती यात्‌रा
- Multiinstitute Nation wide consortium project (IITs, ISI, Jadavpur, Anna University, CDAC)



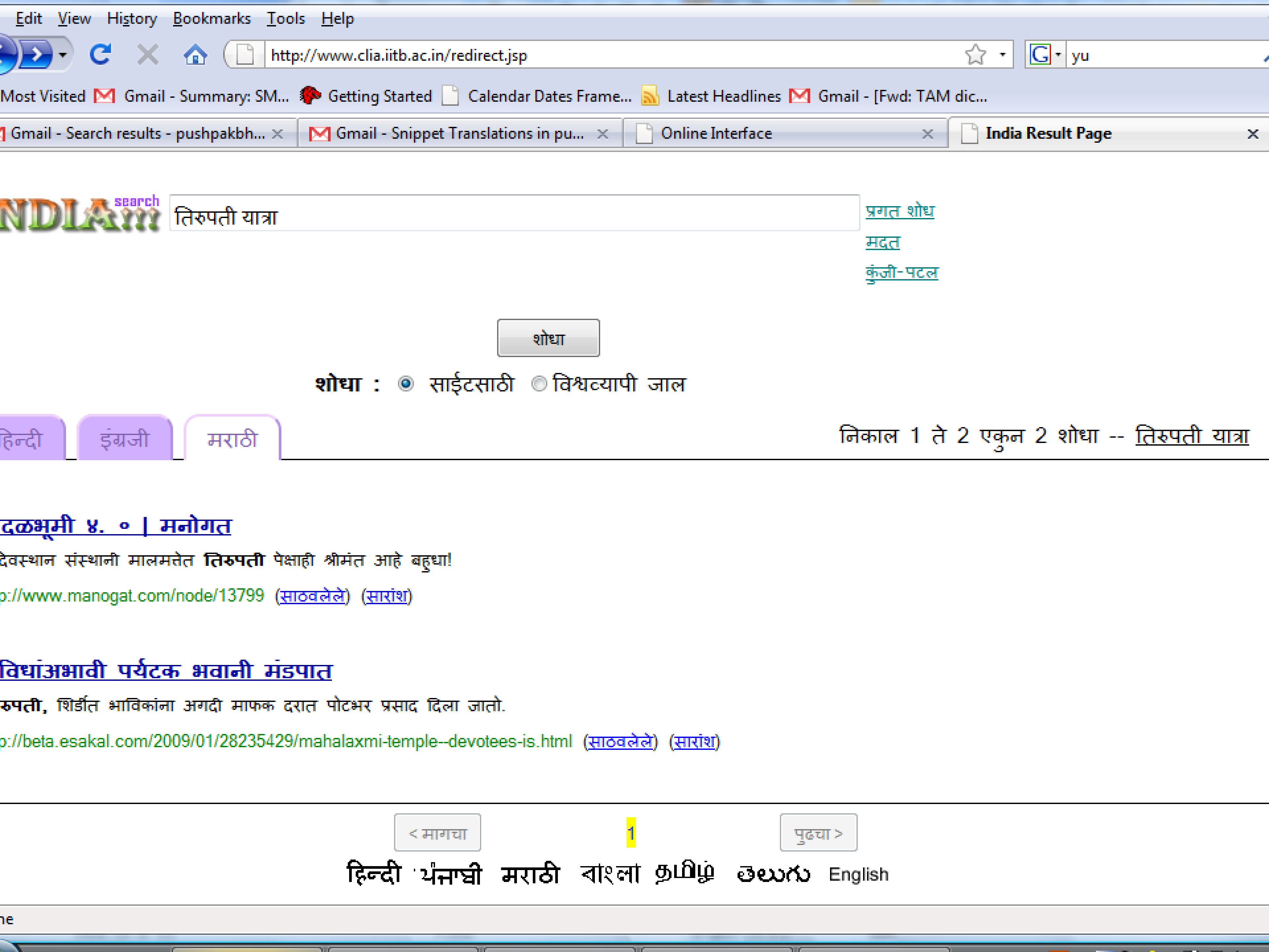
तिरुपती यात्रा

डोमेन: ● पर्यटन आणि तीर्थक्षेत्रे ○ सर्व

शोधा : ● साईटसाठी ○ विश्वव्यापी जाल

इस पेज को अपना मुख्य पृष्ठ बनाएं !

Project Funded by - TDIL, Department of IT, Government of India



http://www.clia.iitb.ac.in/redirect.jsp

yu

तिरुपती यात्रा

शोधा : ● साईटसाठी ○ विश्वव्यापी जाल

निकाल 1 ते 2 एकुन 2 शोधा -- तिरुपती यात्रा

दळभूमी ४. ० | मनोगत

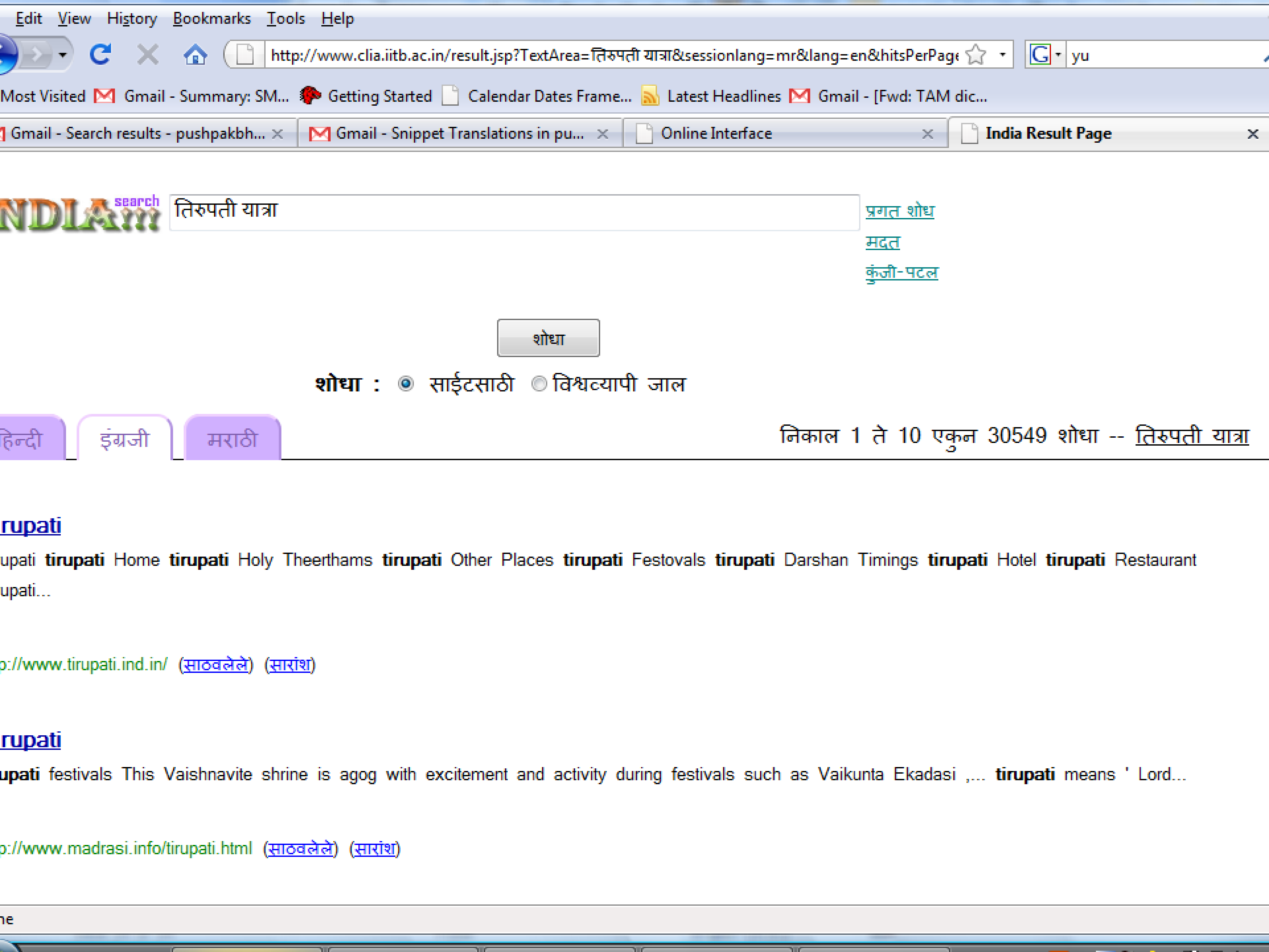
देवस्थान संस्थानी मालमत्तेत **तिरुपती** पेक्षाही श्रीमंत आहे बहुधा!

p://www.manogat.com/node/13799 (साठवलेले) (सारांश)

विधांअभावी पर्यटक भवानी मंडपात

रुपती, शिर्डीत भाविकांना अगदी माफक दरात पोटभर प्रसाद दिला जातो.

p://beta.esakal.com/2009/01/28235429/mahalaxmi-temple--devotees-is.html (साठवलेले) (सारांश)



http://www.clia.iitb.ac.in/result.jsp?TextArea=तिरुपती यात्रा&sessionlang=mr&lang=en&hitsPerPage

तिरुपती यात्रा

शोधा : ● साईटसाठी ○ विश्वव्यापी जाल

निकाल 1 ते 10 एकुन 30549 शोधा -- तिरुपती यात्रा

rupati

upati **tirupati** Home **tirupati** Holy Theerthams **tirupati** Other Places **tirupati** Festovals **tirupati** Darshan Timings **tirupati** Hotel **tirupati** Restaurant upati...

p://www.tirupati.ind.in/ (साठवलेले) (सारांश)

rupati

upati festivals This Vaishnavite shrine is agog with excitement and activity during festivals such as Vaikunta Ekadasi ,... **tirupati** means ' Lord...

p://www.madrasi.info/tirupati.html (साठवलेले) (सारांश)