

---



LING 573  
SLIDES # 1  
April 30, 2007

# PART-OF-SPEECH TAGGER INDUCTION

*Data Cleanup & Tag Projection*

Sabrina Burleigh & Ankit Srivastava

---

# OVERVIEW

M  
I  
L  
E  
S  
T  
O  
N  
E  
S

- ❑ IGT Data Collection in **Hindi** and **Spanish**
- ❑ Cleaning & Filtering IGT Repository
- ❑ Data Standardization
- ❑ Extract, Stem, & Tag English Translation Lines
- ❑ Project tags from Translation to Source Lines via Gloss Lines
- ❑ Lexicon Induction & Analyzer
- ❑ HMM Tagger Induction
- ❑ Evaluation

# WHAT WE HAVE DONE

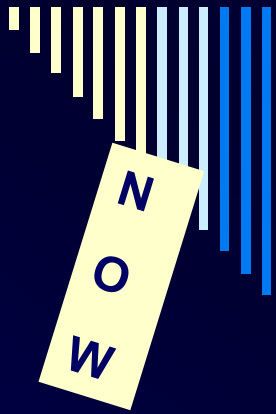
C  
O  
M  
P  
L  
E  
T  
E

- ❑ Crawled the Web for PDF documents containing IGT
- ❑ Discard non Hindi and non Spanish IGT
- ❑ Discard non IGT lines labeled as IGT
- ❑ Combine with IGT from ODIN
- ❑ Split the collection into “3 line” and “Incomplete”
- ❑ Use all the IGTs classified as “3 line” in the next phase

# EVALUATION - 1

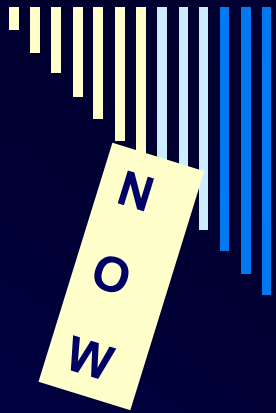
C  
O  
M  
P  
L  
E  
T  
E

	HD LANG
# (Total Docs)	1,947
# (Docs w/ IGT)	83
# (3 line IGT)	606
# (Incomplete)	224






# WHERE WE ARE AT

- Currently organizing projected data
- Processing Translation Line
- Processing Gloss Line
- Processing Source Line
- Investigating Character Discrepancies



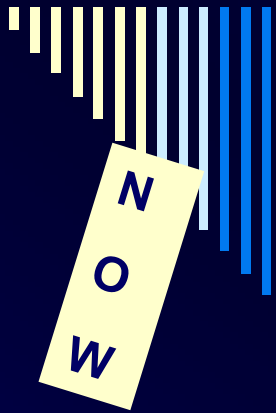
# TRANSLATION

- Translation Line [ 3<sup>rd</sup> Line of IGT ] 
- Ratnaparkhi's Tagger [ /opt/JMX ] 
- English Morpher [ eng\_morph.pl ] 

'Sita thinks that the girl who is singing on TV is beautiful .

'Sita\_NNP thinks\_VBZ that\_IN the\_DT girl\_NN who\_WP is\_VBZ singing\_VBG  
on\_IN TV\_NN is\_VBZ beautiful\_JJ .\_.

'Sita\_NNP think\_VBZ that\_IN the\_DT girl\_NN who\_WP be\_VBZ sing\_VBG  
on\_IN TV\_NN be\_VBZ beautiful\_JJ .\_.



# TRANSLATION - PROBLEM

- Discrepancies in the tagger and stemmer when the same sentence occurs more than once.

**Anu\_.** saw\_VBD the\_DT moon\_NN .\_.

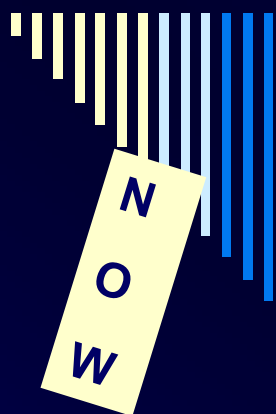
**Anu\_NN** saw\_VBD the\_DT moon\_NN .\_.

`Raam sneezed (volitionally)'

Raam\_NN **sneez\_VBD** (volitionally)'\_.

`Raam sneezed'

`Raam\_NN **sneezed'**.\_.



# PROJECTION - GLOSS

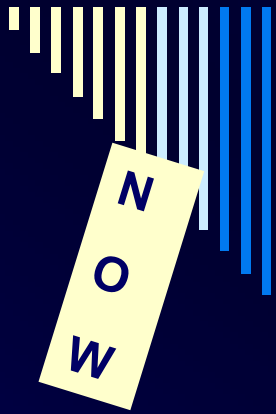
- Project tags from Translation line to Gloss line

(2) Anuu-ko caand dikhii.  
Anu-dat moon-nom appear-perf  
Anu\_NN see\_VBD the\_DT moon\_NN .\_.

Anu-dat

□ Anu\_NN

(2) Anuu-ko caand dikhii.  
Anu-dat moon-nom\_NN appear-perf  
Anu\_NN see\_VBD the\_DT moon\_NN .\_.

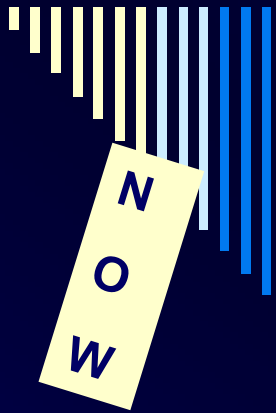


# PROJECTION - SOURCE

- Project tags from Gloss line to Source line

(2) Anuu-ko caand dikhii.  
Anu-dat moon-nom\_NN appear-perf  
Anu\_NN see\_VBD the\_DT moon\_NN .\_.

(2)/Anu-dat Anuuko/NN caand/appear-perf  
Anu-dat moon-nom\_NN appear-perf  
Anu\_NN see\_VBD the\_DT moon\_NN .\_.



# EXAMPLE FROM HINDI

- Where Stemmer does not help
- Stem the gloss line also?

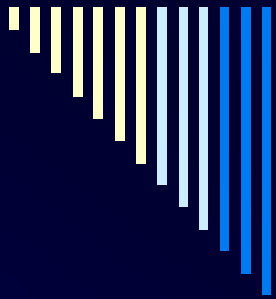
Siitaane/NNP Hari-ko/NNP kitaab/NN khariid-ne-ko/VB kahaa/told  
Sita-erg\_NNP Hari-dat\_NNP book\_NN buy-inf\_VB told  
'Sita\_NNP tell\_VBD Hari\_NNP to\_TO buy\_VB a/the\_NN book\_NN .\_.



# OTHER PROBLEMS FACED

S  
T  
R  
I  
N  
G  
  
E  
D  
I  
T

- Hindi Transliteration (Vowels, Retroflex, Nasals)
- Accented characters in Spanish disappear
- Multiple tags for the same word
- Data needs further cleaning



# WHAT IS LEFT

- Morphological / Distributional Analyzer
- Lexicon [ IGT and Raw Text ]
- HMM Tagger