Politeness Recognition Tool (PoRT) for Hindi

Ritesh Kumar, Centre for Linguistics, riteshkrjnu@gmail.com
Supervised by: Dr. Girish Nath Jha (Special Centre for Sanskrit Studies) & Dr. Ayesha Kidwai (Centre for Linguistics)
Jawaharlal Nehru University, New Delhi

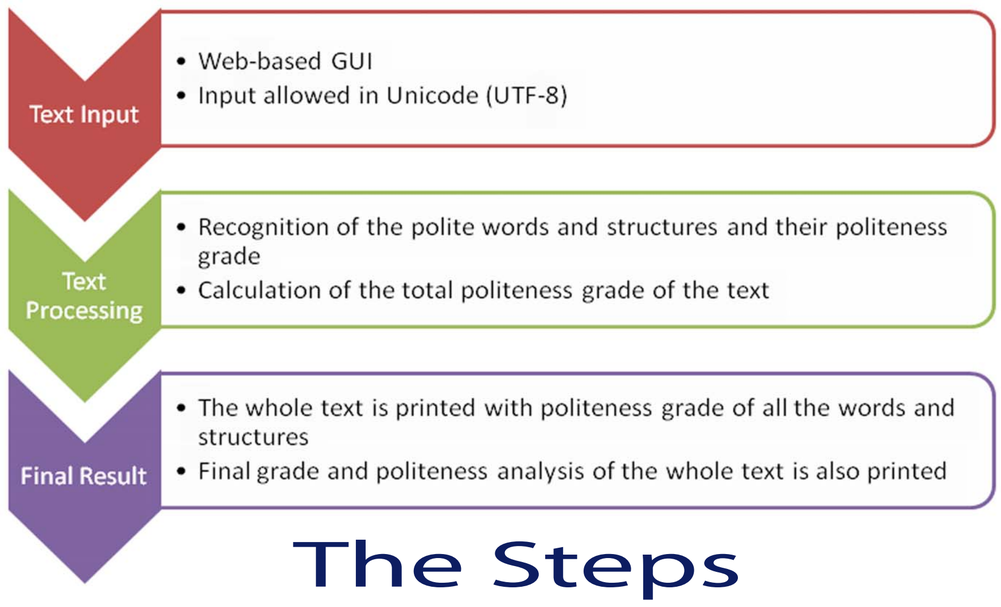
Description

Linguistic politeness refers to the usage of language in such a way that it the equilibrium in the personal relationships is not disturbed and the communication takes place effectively. This usage of language is manifested in different kinds of structural and lexical choices (which language and culture-specific) that a user of the language makes.

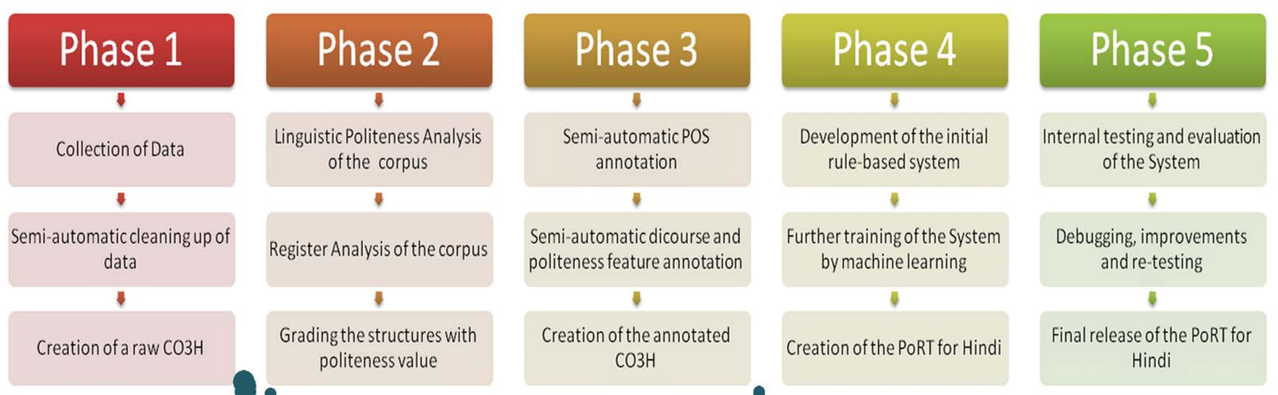
We are working towards develping a computational tool which could determine the extent to which a given text in Hindi is polite or impolite. For this purpose we have prepared a corpus of computer-mediated communication in Hindi (CO3H). It will be used for both the linguistic analysis of linguistic politeness in Hindi and the training and testing of the system. Selection of CMC is a conscious decision to eliminate the prosodic aspects of the language that may lead to (im)politeness. Since the system will take the input in the form of written texts, training and testing of the system on the data that is closest to spoken data in the written form would be the ideal position.

This poster gives a broad schema of the system and the developments that have been made till now.

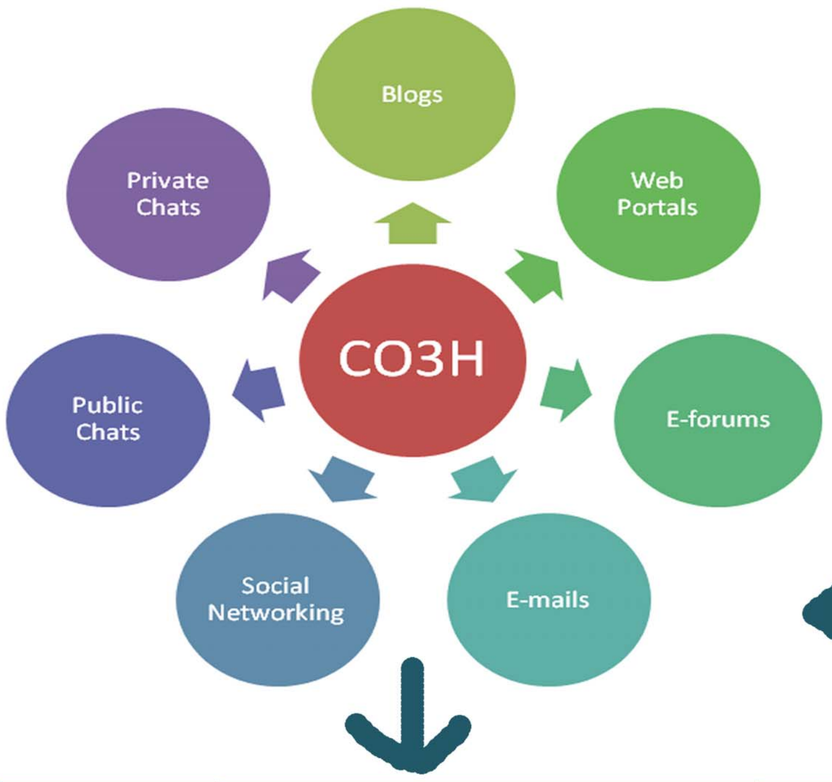
The Goal



The Steps

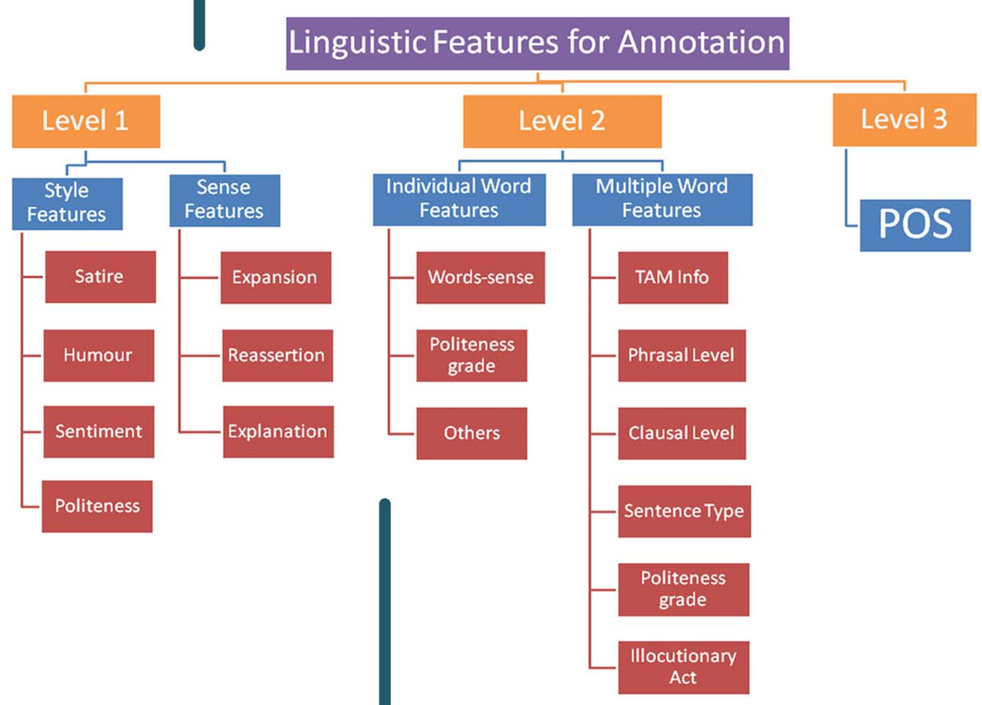


The Corpus



Media	Entries/Transcripts Collected	Immediate Target
Blogs	100 blog sites, totaling more than 2500 blog entries	At least 5000 blog entries
Web Portals	2 web portals, totaling around 200 entries.	At least 8 different web portals
E-forums	5 Google groups, totaling around 500 discussions.	At least 200 Google groups
E-mails	> 10,000 e-mails but only few hundred among them are in Hindi	At least 5000 Hindi e-mails
Public Chats	Public chats of around 150 days	Till 31st December, 2010
Private Chats	More than 1000 private chat transcripts	At least 2500 chat transcripts

Linguistic Features for Annotation



The Progress