

A Register-based Annotation Scheme for CO3H

Ritesh Kumar
Centre for Linguistics
Jawaharlal Nehru University
New Delhi, India
riteshkrjnu@gmail.com

ABSTRACT

This paper gives a description of an annotation scheme for annotating a corpus of computer-mediated communication in Hindi (CO3H) with certain semantic, pragmatic and situational features. The annotation scheme is based on the theory of register analysis, where it is assumed that a registeral difference entails difference in certain linguistic features. It adapts and integrates the annotation schemes of sense annotation in the Penn Discourse Treebank and dialogue act annotation of DIT++ within this larger registeral framework. The situational and linguistic features that will be used to annotate the corpus for PoRT is described in the paper, along with some proposed labels for these features.

Categories and Subject Descriptors

I.2.3 [Artificial Intelligence]: Natural Language Processing—*Discourse, Text analysis, Language parsing and understanding*

General Terms

Theory, Design

Keywords

Annotation, register analysis, CMC annotation, CO3H, PoRT

1. INTRODUCTION

Semantically and pragmatically annotated corpus is one of the primary resources for developing any semantics-based tool or application. Recent attention to this task have resulted in the development of quite a few semantically annotated corpus like Penn Discourse Treebank (for a more detailed overview refer to [5]). However for a computationally less-studied language like Hindi, such a resource does not exist as of now.

In the present paper a model to annotate a corpus of computer-mediated communication in Hindi (CO3H) with

semantic and pragmatic features is described. The annotation will be primarily done to develop a politeness recognition tool (PoRT) for Hindi (given a text this tool will recognise if the text, which may be one word, sentence, paragraph or a whole book, is polite or not and to what extent). However, the scheme will be developed in such a way that it could be later expanded and adapted for other purposes also. The scheme is being developed keeping register analysis as the primary theoretical background. This attempt is novel in three principal ways—

- There is no previous attempt to annotate *Hindi* data with semantic and pragmatic features.
- As far as I know, we do not have any corpus computer-mediated communication (CMC) annotated with these informations, although there have been quite a few attempts to annotate spoken, face-to-face dialogues and multimodal data [7, 4, 6]. Consequently a new approach to annotate CMC with semantic and pragmatic information needs to be devised.
- There is no previous attempt to develop an annotation scheme based on register analysis.

The creation of this resource involve the following steps:

1. Collection, cleaning up and arranging the data from CMC in a proper format to prepare the raw corpus.
2. Register analysis of the data from different kinds of CMC so as to deduce proper registers and relevant linguistic features for automatic classification.
3. Politeness analysis of the data for proper annotation of politeness levels. Since the annotated data at this point will be used for the training of PoRT, it is necessary that the data is annotated with the information regarding how polite a particular word, structure or whole discourse is. Furthermore it needs to be generalised up to the level where the machine could automatically classify texts at the level of politeness.
4. Development of the tagset for annotation beyond the POS level (tagset for POS annotation is already built), for annotating with some of the semantic and pragmatic properties.
5. Automatic annotation of each word with the POS tags. It will be done using an annotation tool trained on 50,000 manually annotated sentences from the Hindi corpora of Indian Languages Corpora Initiative (ILCI).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS '11, May 25-27, 2011 Sogndal, Norway

Copyright 2011 ACM 978-1-4503-0148-0/11/05 ...\$10.00.

The annotation will be done using the BIS tagset, which is the national standard for POS annotation of Indian languages.

6. A semi-supervised automatic annotation of the data at different levels beyond the POS. An initial rule-based tool will be developed for the annotation. The data annotated by this tool will be verified, modified and corrected by the human annotators and the final annotated data will be used for training the annotation tool.

Till now we have made considerable progress on 1, 2 and 3. In this paper we present a short summary of the work done so far and the initial work towards achieving the fourth step.

2. THE CORPUS

The corpus consists of data taken from six different kinds of computer-mediated communication. This division is made on the basis of the kind of media that is being used to carry out the conversation. They include:

1. **Blogs:** Data for the blogs is collected using

<http://chitthajagat.in> [a Hindi blog aggregator]. Till now Data is collected from around 100 blog sites, totaling more than 2500 blog entries. The collection of data is an ongoing job but the initial target is to collect at least 5000 blog entries that have been written in between 1st January, 2007 and 31st December, 2010. The conversation in a blog mainly involve the discussions, replies and arguments in the comments section. The main blog entry generally forms the broader thematic context of the discourse. However sometimes some fictional conversation or conversation in the form of interviews or third narration of conversation could also be found. All of these are not always computer-mediated communication; however the interviews conducted online are definitely CMC.

2. **Web Portals:** Data for web portals is collected using eight different web portals in Hindi. Web portals are very similar to the blogs in the sense that here also the conversation takes place mostly in the comments section. At the surface level both of them look very similar but when we carry out a detailed register analysis we find quite significant systematic linguistic differences between the two ([2]). Till now data for portals is collected from two web portals, totaling around 200 entries. The target is to collect data from at least eight different web portals that are written in between 1st January, 2007 and 31st December, 2010.

3. **E-forums:** Data for e-forums is collected using Google and Yahoo groups in Hindi. Till now Data for e-forums is taken from five Google groups in Hindi, totaling around 500 discussions. E-forums are by its very nature meant for discussions. So as soon as there is a reply to one of the threads started by anyone then it is assumed that some kind of communication has taken place. So all those forum entries which have received at least one answer are included in the corpus. The initial plan is to collect data from at least 200 Google groups (as per the data on Google group directory, there is a

little over 1000 Hindi groups) in which the conversation has been initiated in between 1st January, 2007 and 31st December, 2010.

4. **E-mails:** Data for e-mails will be taken from the e-mail accounts of 8 different people (with their proper consent and their being fully aware and having a good understanding of what exactly the data will be used for). Till now 3 people have already given access to their e-mails, which count to more than 10,000 e-mails but only few hundreds among them are in Hindi. Five other people have agreed to give access and soon they will also be acquired. All the e-mails that will be included in the corpus will be written in between 1st January, 2007 and 31st December, 2010.

5. **Chats:** Data for public chats is taken from the log files of IRC chat on the channel 'India' of 'Dalnet' network. Public chats of around 150 days have been logged but again, as in e-mails, chats in Hindi will need to be sieved out of them. The log files are taken from 24th April, 2009 and will continue till 31st December, 2010. This is the only exception in the corpus whose data is not available from 1st January, 2007.

For private chats the data from Gmail chat transcripts is provided by the same people who gave it for the e-mails. More than 1000 private chat transcripts in Hindi is acquired (average length of transcripts is around 80-100 lines, with transcripts being as small as two lines and as big as 500 or more lines).

6. **Social Networking Sites:** Data from social networking will be collected from Facebook and Orkut. The conversation on the social networking sites include the replies to the comments posted on the photographs, one's status, the wall postings and the conversation through messages (on Facebook) and the scraps (on Orkut). There has to be at least 3 comments (as in blogs and web portals) or at least one reply to a message on the facebook or at least 3 replies to a scrap on orkut for it to be included in the corpus. And as in other cases all these must have taken place in between 1st January, 2007 and 31st December, 2010. The data collection for social networking sites have not yet started but we will start it very soon.

The data is then pre-processed (to clean up all the noise), arranged and annotated at different levels. Annotation at the POS level will be done using the BIS tagset which is now the Indian standard for POS tagging approved by the Bureau of Indian Standards.

After POS tagging, semantic/pragmatic level tagging of the data will be done. The present paper gives a detailed description of the proposed annotation scheme and a basic tentative tagset (which will be expanded as the manual tagging will begin).

3. THE THEORETICAL MODEL

The corpus will be tagged using an annotation scheme based on the theoretical assumptions and bases of register analysis (integrated with the dialog act annotation scheme <http://dit.uvt.nl/>). Registers are generally defined as the varieties of a language that is characterised by distinct linguistic features across different situations of use. The

model for register analysis that is closely followed in the development of this annotation scheme is discussed in great details in [1].

Register analysis mainly consists of two components:

1. The first is the analysis and description of the *situational context*. A particular situational context is made up of different kinds of characteristics and broadly they are described in the following terms:
 - (a) Description of the addressee and the addressor (called *participants* in the conversation) in terms of their social characteristics like age, education, etc and whether there are one or more addressor and addressee.
 - (b) Description of the relation among participants.
 - (c) Description of the channel of communication (which include both the *mode*, like, writing, speaking, etc. and the *medium* of the communication, like telephonic, computer-mediated, face-to-face, letters, etc.)
 - (d) Description of the *production circumstances* of the communication. It will include information like whether the conversation is planned or real time and whether some revision and editing is being done before anything is communicated to the other person.
 - (e) Description of the fact that whether the place of communication is public or private and whether time and place is shared by all the participants.
 - (f) Description of the *communicative purpose* of the conversation
 - (g) Finally the *domain or topic* of the conversation

A description of all these characteristics is very necessary since they give a clue to the kind of language and the linguistic features that we may encounter. Also it must be noted that a lot of times one of these features entail some of the other features and a separate description of each feature may not be required all the time.

Let us take the example of the situational analysis of different kinds of CMC. CMC implies that the mode of communication will be always writing (it is true in all situations, except video/voice chat, which is anyway not included in the present corpus); however there are six different kinds of media (in the present corpus). Moreover, if the media of CMC is not chat then the text produced could be assumed to be revised and edited (although this necessarily may not be the fact but the media provides this option to the user). Similarly the place of communication is private (except in public chat) and the time and place is not shared among the participants (in some cases time might be shared). However the communicative purpose and the topic have a great possibility to differ across different communications and so they need to be analysed each time. Moreover the description about the participants is not expected to differ much in some aspects (like the users of CMC are generally young/middle-aged and educated in the current scenario) but in some aspects it

varies a lot (as in the gender and profession of the participants). Similarly, depending on the kind of media one is using, relation among participants is more or less the same.

Despite these generalizations, one thing that is absolutely essential is that a thorough analysis is carried out every time because these generalizations may actually be the result of stereotyping and common-sense understanding. Moreover as we see there are very few absolute generalizations and it is very necessary to describe those situations also.

2. The second step in register analysis is the analysis and description of the unique as well as non-unique linguistic features related to the words and the structures that are commonly used in one particular situational context. This step has unlimited possibilities and any linguistic feature under the sun could be analysed in this step (for a highly detailed list of possibly relevant list of features that could be analysed, see [1]).

Register analysis is chosen as the theoretical basis for the annotation scheme here since it provides an excellent ground for the text classification and at the same time allows for a very rigorous and minute linguistic analysis of the text. In fact, the classification of text in the register analysis is based on the linguistic features.

4. THE ANNOTATION SCHEME

For the purpose of annotation, we have taken only two situational characteristics under consideration:

- The communicative purpose
- The topic of communication

The other information is not being annotated for the following reasons:

- Some of the information like those of media (and sub-media) of communication are included elsewhere, as in the metadata.
- Lack of certain information like those about participants and the relation among participants.
- Giving all the details would make the annotation scheme very complex reducing the performance of both the human annotators and machine, which may ultimately defeat the purpose of this classification.
- Some of these information may not be actually essential for the present purposes.

However the annotation scheme is still not freezed and so if the need arises then the annotation scheme could include more information.

These situational characteristics, followed by the linguistic characteristics have the potential to form a complete annotation scheme. The linguistic features could include anything from sense of the word to sentence type to part of speech.

In this scheme, all these features are arranged hierarchically, with the situational characteristics forming the top-level of the hierarchy and the linguistic features coming at the lower level (POS being the lowest and the most minute

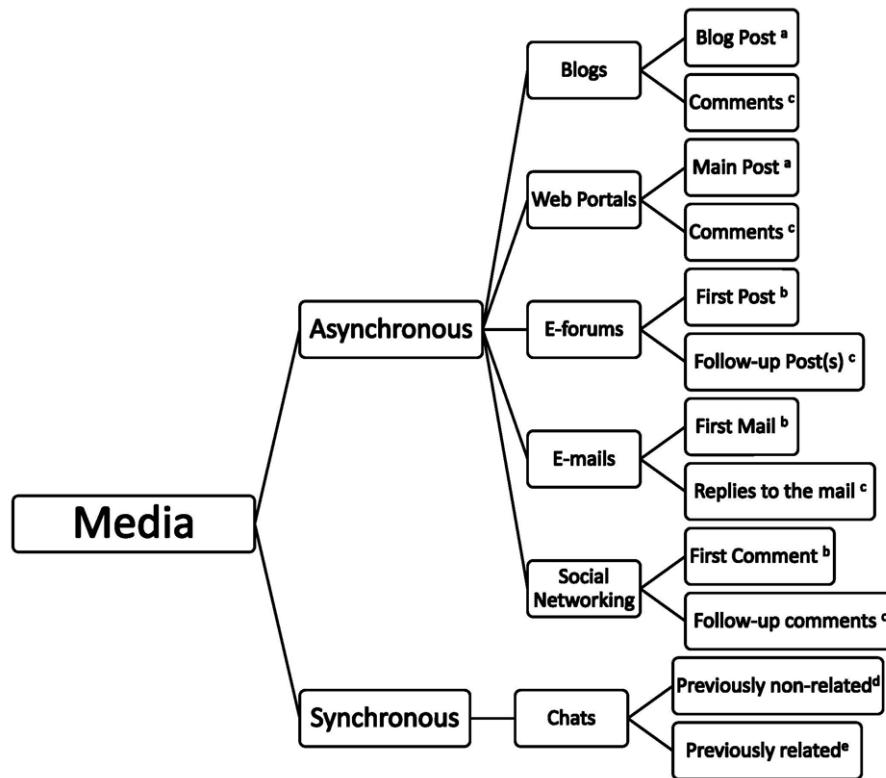


Figure 1: Situational Characteristics related to the media for communication

level in the category). A description of the situational features which is relevant at the discourse level and also play very significant role in shaping the linguistic features and structures in discourse (hierarchical representation shown in Fig. 1 and Fig. 2) are given below. It is to be noted that these situational features characterise one dialog act, which may range from one word (as in greetings) to one sentence (as in some of the comments or in chats) to one or multiple paragraphs (as in the whole of blog entry).

1. Level A: The main posts on a blog or web portal are grouped under type A.

(a) Opinions: The blog posts or posts on the web portals which express opinion regarding some issue have this characteristic

- i. Socio-political Issues: The blogs posts which express opinion on some socio-political issues like rise of naxalism, women's position in the society, etc have this characteristic.
- ii. Movie Reviews: The blog posts which review full-length feature films, short films, documentaries, TV series or any kind of video have this characteristic.
- iii. Other Issues: The blog posts about any other issues, like a travel blog, food review or anything else have this characteristic. This could be later expanded into other individual characteristics, as more data is analysed.

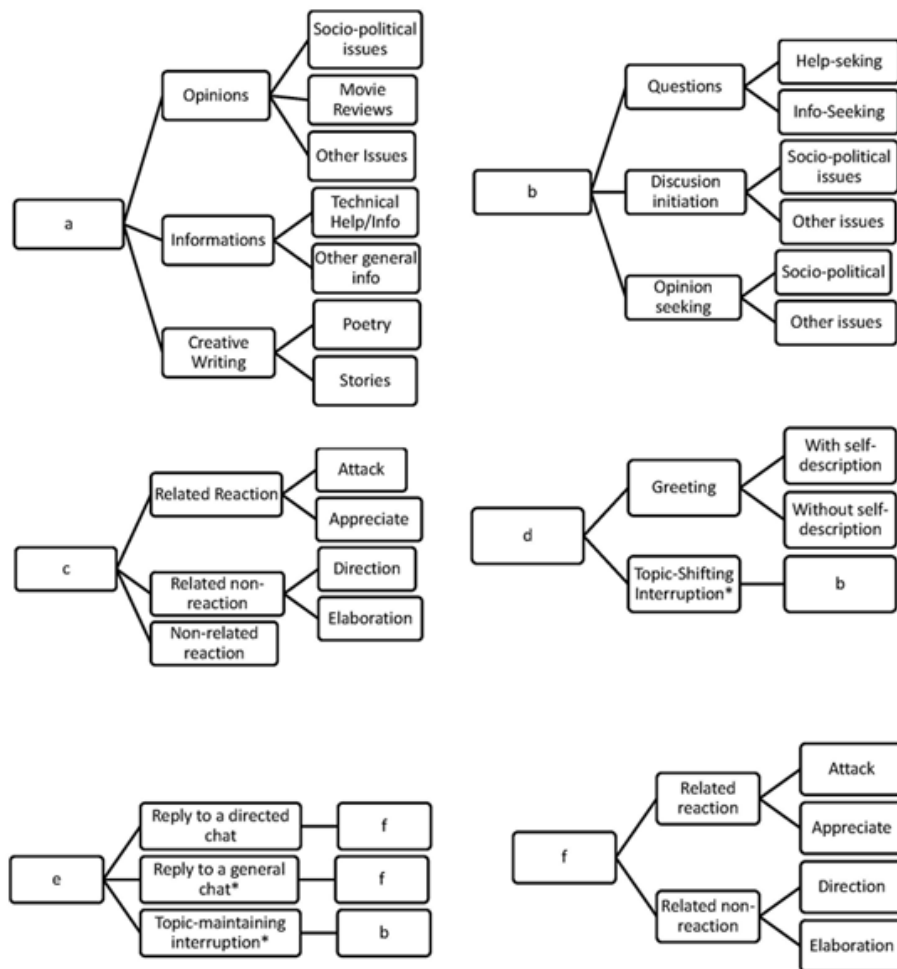
(b) Informations

- i. Technical Help/ information: The posts on blogs or portals which give information or some kind of help materials related to the technical aspects (especially with respect to the use of latest technologies in the field of computer science, mobile technologies and other similar technologies).

- ii. Other General information: The blog posts which give any other kind of information (other than the technical ones) could be given this characteristic. It must be kept in mind that both this and 'other issues' in opinions could have the same theme. Take for example, if a travel blog is giving the views regarding what place to go and how good the place is then it has the characteristics of an opinion; on the other hand, if it just describes the place, how to reach there and what all is there to do then it should come under informations. At times there is a very thin line between the two. At such places the decision of the annotator should be guided by what (s)he thinks to be the primary motive of the writer (whether it is to 'inform' about some place or it is to 'give opinion' regarding the place).

(c) Creative Writing: The original creative, fictional works in the blogs or the portals have this characteristic.

- i. Poetry: Anything which is not written in prose will be included here.



*Not applicable for Private Chats

Figure 2: Situational Characteristics related to each media

- ii. Stories: Everything that is written in prose is included here.
 2. Type C: All the comments that have been given to the main blog post/ portals/ replies to the e-mails/ follow-up comments on the e-forums and social networkign sites are covered under this characteristic.
 - (a) Related reaction: The comments which are related to the main/ first post in some way or the other and it reacts to what has been said there are included here.
 - i. Attack: The comments which attack and counter the main post are included here.
 - ii. Appreciate: The comments which appreciate whatever is written in the main post are included here.
 - (b) Related non-reaction: These comments are also related to what is said in the main/ first entry but they do not react to what is said
 - i. Direction: These comments tells the writer of the post regarding some other similar works and how to access them; the writers are 'di-rected' to do something (and this is in relation to what is written in the blog entry).
 - ii. Elaboration: These comments further elaborate on whatever is said in the post; they either elaborate further or give some new in-formation regarding the same issue.
 - (c) Non-related reaction: Sometimes it is noted that the reactions of the commentators are not directed towards the content of the post; rather it is to-wards other things like the language or some other issue. Then such comments will be included here.
 3. Type B: All the texts taken from the first post on an e-forum/ first e-mail/ first comment on a social net-working site/ topic-shifting or topic-maintaining inter-ruption on a public chat have this situational char-acteristic. In public chat, when two are more people are chatting on some theme and then somebody inter-rupts by sending a chat which is completely unrelated to what the people are chatting about then it is termed as the "topic-shifting interruption". On the other hand, when somebody Replies to a chat which is not directed towards the one who is replying but towards someone else then it is termed as the "topic-maintaining inter-ruption". In contrast to the topic-shifting interruption, here the reply is on the same theme and related to the previous chat. The purpose and the topics in these types of situations include
 - (a) Questions: If some question is asked regarding some issue then it is included here.
 - i. Help seeking: The question is meant to seek help in operating and using some gadgets, softwares, etc.
 - ii. Information seeking: The question is related to seeking information on 'how good' or 'how bad' something is and where to get it, etc.
 - (b) Discussion initiation: If no question is asked but some comments are given on some issue then it is included here.
 - i. Socio-political issues: If the issue is socio-political then it is included here.
 - ii. Other issues: If the issue is something other than socio-political then it is included here. This category has a scope of further expansion after further analysis of the data.
 - (c) Opinion seeking: If somebody is seeking the opin-ion of other people on some issue like, the recent commonwealth games in India, then it is included here. The opinion-seeking has to be completely explicit to be included here.
 - i. Socio-political opinion: If the opinion is sought on some socio-political issue then that is in-cluded here.
 - ii. Other opinions: If the opinion is sought on some other issue then it is included here. This category can, again, be expanded depending on further analysis.
 4. Type D and E: These are very restricted situational characteristics that are found generally only in chats (the major reason behind this difference from the other CMC seems to be the synchronicity of the media). Type D refers to the 'non-related chats'. They include the beginning of a chat (in private chats) or anything which is sent as chat and which is not a reply but an initiation (in public chats). They include only two purposes:
 - (a) Greeting.
 - Introductory Greeting
 - With self-description
 - Without self-description
 - Valedictory Greeting
 - (b) Theme-shifting interruption (only for public chats): Described above
- Type E refers to the 'related chats'. They are the chats that are sent as replies to something which has been already said (that could be non-related or related chat). The purpose of related chats are as follows
- (a) Reply: Those related chats which are reply to something that has been said earlier is included here.
 - i. Reply to a directed chat: Reply to a chat which is directed towards the person who is replying (it is always so in private chat but different in public chat).
 - ii. Reply to a general chat (only for public chats): Reply to a chat which is not directed towards any one individual in particular but towards anyone in the room who wants to reply is in-cluded here.
 - (b) Theme-maintaining interruption (only for public chats): Discussed earlier.
- Moreover these features are divided into two broad levels for the purpose of annotation - Level A and Level B. Any dialog act could be annotated with one feature from each level.

With each of these situational characteristics, a large number of linguistic features could be associated. It must be noted that the set of linguistic features associated with one situational characteristic need not be unique with respect to the linguistic features of other situational characteristics. Thus any number of linguistic features could be shared among the texts having different situational characteristics.

These linguistic features will also be arranged in the hierarchical manner among themselves (Fig. 3). A broad outline of these linguistic features is given below.

1. Level 1: Inter-sentential/discourse level features – The linguistic features which could arise out of the interaction of two or more sentences (in certain instances even a single sentence could depict these features). It could involve a lot of features, some of which are described below:
 - (a) Style features: These features are related to the style in which the text is written. One of the most important thing about these features is that they are not deducible from the linguistic features (nor are they part of the situational characteristics). These features come out by taking into account the pragmatics of the sentence (which is a result of the interaction of situational characteristics and the semantics of the sentence). Some of the categories within style features include:
 - i. Satire/Irony: It involves an attack on someone or something in indirect way. Just one sentence can never make it clear whether it is a satire/irony or a genuine sentence. It is very difficult to recognise such instances automatically. Let us take the following example from the corpus-
 - a) [Rajatji! Raat ko kuchh blue film type ka bhi masala dikhao. TRP danadan badegi.]/ST1 (Rajatji! Kindly show some spicy items like blue film. TRP will soar very quickly). This sentence is written as a reply to the the owner of a news channel who is trying to justify the presentation of false, frivolous and filthy news on his channel in the name of business and profit.
 - ii. Humour: For any sentence to be humorous, the situational characteristics have to be taken into account. Then the pragmatics of the sentence create humour. Humour could be further sub-divided into ‘satirical humour’ (as is created in example (a) above) and ‘genuine humour’.
 - iii. Sentiment expression: The same sentence may express different kinds of sentiments (generally a distinction between ‘positive’ and ‘negative’ sentiment is being made in the literature on ‘sentiment analysis’) depending on the situational characteristics. For example, let us take the example above. It depicts ‘negative’ sentiment, which is also derived from the satire. However the following example shows a positive sentiment.
 - b) [Jaan kar achha laga ki tehelak men aise bhi lekh hain jo ki life vritant hain. ab se

main regular is patrika ko dekhunga.....]/ETP (It was good to know that there are also such articles in Tehelka which describes life. From now on I will regularly read this magazine) Besides these we will also have sentences which do not express opinions as such and so they will be considered ‘neutral’ as far as sentiment is concerned.

- iv. Politeness: Politeness, like other style features, is extremely context-sensitive. The major thrust and motivation behind developing this annotation scheme is to enable machines to recognise politeness automatically. Politeness is an extremely complex and varied phenomenon, which is manifested at and affected at all the three levels of linguistic representation - lexical, syntactic and discourse. Thus we have several lexical items like the words for excuse, thanking, congratulating, etc which may be termed polite and then there are swear words and slangs which are generally impolite. Similarly at the syntactic level some choice of sentence types (like using a question instead of an imperative) and certain agreement features (like agreement of verb with the honorificity of the subject/ object) are considered polite. However it is at the discourse level (when context comes into play) that the situation becomes really tricky and complex. At this level factors like familiarity among the participants, relationship among them and the power relation among them begin to come into play and interfere with the politeness level which is manifested at the previous two levels. For example use of an extremely polite form with a friend is considered a sarcasm/ satire, which is not polite. Moreover as we can see the presence/absence of other style features also affect the politeness value at the discourse level. So a satire or a negative sentiment would invert the politeness value if a text is adjudged polite while humour or a positive sentiment would invert it if the text is adjudged impolite after taking into consideration lexical and syntactic level.
- (b) Inter-sentential sense-related features: Annotation of certain sense is necessary for the understanding of relation among different sentences. Annotation for the labelling of anaphora and labelling of connectives ([3]) are some of the examples of annotation of these features. The linguistic features/functions that could be understood through these annotations include:
 - i. Expansion: When two or more sentences are written to further describe an object, event or anything else then they are connected to each other by the function of expansion. Suppose someone is describing Patna (a city in India) then all the sentences that are used to describe the city one after the other are the expansion of the what is previously said. The same relation also holds in the cases where a concept is defined in one sentence and then

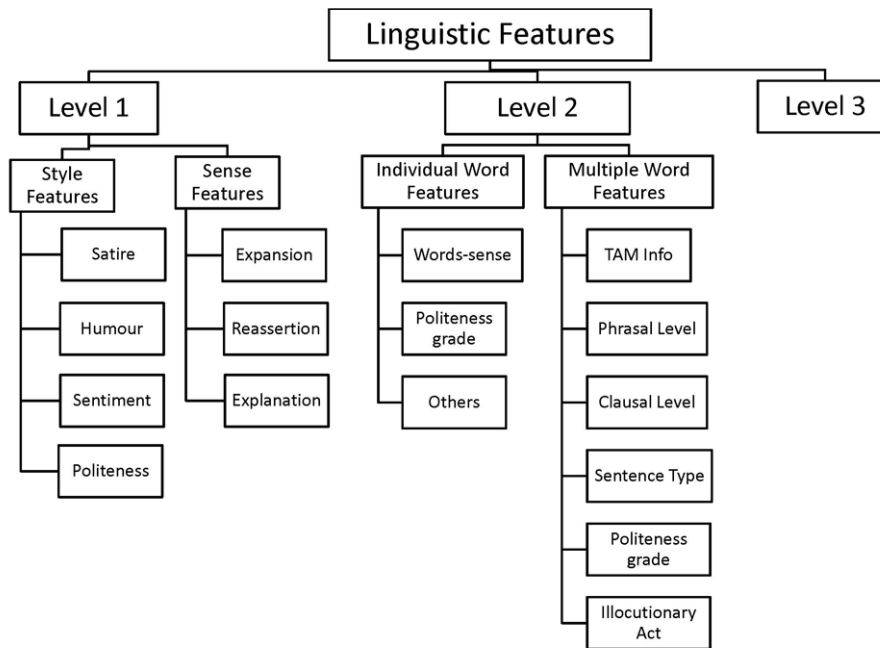


Figure 3: Linguistic Features for Annotation

it is explained in more detail by breaking up into several sentences.

- ii. Reassertion: When the same thing is repeated in more than one sentence then they are performing the function of reassertion. It may or may not involve a change in the grammatical structure of the following sentence. It differs from expansion in the fact that neither any new thing is added in the following sentence nor anything is broken up; rather exactly same thing is restated.
 - iii. Explanation: When the following sentences explain the rationale behind one sentence then they are performing the function of explanation. The following sentences give the reason why that which is said earlier should be done or followed or considered correct.
2. Level 2: Intra-sentential features – The linguistic features at this level cover those aspects of language, which could be described by referring to a single word (beyond POS), a group of words or a single sentence. These features generally work at the sentence level and do not need to go beyond a sentence to be described.
- (a) Individual word features: The features which belong to a single word and do not refer to other words in the sentence for their understanding and description.
 - i. Word-sense features: These refer to the different senses or synonyms that a word may have. It must be noted that the words acquire one particular sense in a given context but the fact that one particular word has 3 or

4 senses is a feature of that word only. For example, in order to know that the word ‘bank’ in English has at least 3 different senses and we know what they are. In order to know this and only this we need not know the context.

- ii. Politeness grade: Each word is assumed to be polite or impolite to certain extent. This feature specifies to which extent is the word (im)polite. At present the words are categorised into five levels of politeness - absolutely polite, somewhat polite, neutral, somewhat impolite and absolutely impolite. Each word is tagged with one of these values of politeness. The labelling of words will be done without taking into consideration the context.
 - iii. Other features: This feature is added to facilitate the expansion of the scheme. In case some other individual word features need to be tagged then those features could be added to the scheme.
- (b) Multiple word features: The features which could be described only when more than one single word is taken into consideration.
- i. Word-group features: Features related to some group of words (like complex verbs, with further sub-division of compound and conjunct verbs in Indian languages) are could be handled at this level.
 - ii. TAM Category: Information like tense, aspect and modality of the sentence could be provided at this level. Generally verb complex (consisting of at least a main verb and an auxiliary verb) carry these information.

Table 1: Situational categories to be annotated for the development of PoRT

Situational Category				Labels
Top Level	Level 1 (Purpose)	Level 2 (Topic/ Domain)	Level 3 (Sub-Topic/ Sub-Domain)	
Level A				
	Greeting			GR
		Introductory		ING
			With self-description	PSI
			Without self-description	NSI
		Valedictory		VAG
	Opinion Expression			OP
		Socio-political issues		SPO
		Movie Reviews		MRO
		Other issues		OIO
	Information Providing			IP
		Technical help/info		THI
		Other general info		OGI
	Opinion seeking			OS
		Socio-political issues		SSO
		Other issues		OSO
	Information seeking			IS
		Technical Help/ Info seeking		HSI
		Other General Help/ Info seeking		ISI
	Discussion initiation			DI
		Socio-political issues		SPD
		Other issues		OID
	Creative Writing			RC
		Poetry		PR
		Stories		SR
	Related reaction			RR
		Attack		ATR
		Appreciate		APR
	Related non-reaction			RN
		Direction		DIR
		Elaboration		ELR
	Non-related reaction			NC
Level B				
	Initiation			ID
	Interruption			IT
		Theme-Shifting		TSI
		Theme-maintaining		TMI
	Reaction/ Reply			RD

Table 2: Linguistic features to be annotated for the development of PoRT

Linguistic Category					Labels
Top level	Level 1	Level 2	Level 3	Level 4	
Inter-sentential features					IE
	Style Features				SF
		Satire			ST
			Yes		ST1
			No		ST0
		Humour			HT
			Yes		HT1
			No		HT0
		Sentiment			ET
			Positive		ETP
			Neutral		ETU
			Negative		ETN
		Politeness			PT
			Absolutely Polite		APT
			Somewhat Polite		SPT
			Neutral		NPT
			Somewhat Impolite		SIT
			Absolutely Impolite		AIT
	Sense Features				EI
		Expansion			EE
		Reassertion			RE
		Explanation			XE
Intra-Sentential features					IA
	Individual Word Features				II
		Word Sense			WI
		Politeness Grade			PI
			Absolutely Polite		API
			Somewhat Polite		SPI
			Neutral		NPI
			Somewhat Impolite		SII
			Absolutely Impolite		AII
	Multiple Word Features				MI
		Tense Info			TM
			Present		PRT
			Past		PST
			Future		FUT
		Aspect Info			AM
			Simple		SIA
			Progressive		PRA
			Purposive		PUA
		Modality Info			MM
			Declarative		DCM
			Subjunctive		SJM
			Conditional		CDM
			Imperative		IPM
			Presumptive		PSM
			Habitual		HBM
			Abilitative		ATM
		Sentence Type			SM
			Affirmative		AFS
			Negative		NGS
			Interrogative		ITS

Continued in Table 3

Table 3: Linguistic features to be annotated for the development of PoRT: Continued from Table 2

Top level	Level 1	Level 2	Level 3	Level 4	
		Illocutionary Act			IM
			Directive		RI
				Directional	DIR
				Indirectional	INR
				Suggestive	SUR
				Conditional	COR
				Negative Con- ditional	NCR
				Explanative	EXR
				Dissociative	DSR
				Possibilative	PSR
			Assertive		AI
			Commissive		CI
			Expressive		XI
			Declarative		DI
		Sentential Polite- ness grade			PM
			Absolutely Polite		APM
			Somewhat Polite		SPM
			Neutral		NPM
			Somewhat Impo- lite		SIM
			Absolutely Impo- lite		AIM
POS features					As per BIS tagset

iii. Illocutionary Act: The illocutionary acts are the same described by [8].

- A. Assertive: It involves such acts as boasting, complain, deducing, describing, etc.
- B. Directive: It involves such acts as asking, ordering, requesting, inviting, advising etc. It could be further divided into the following sub-categories.
- C. Commissive: It involves the act of promising.
- D. Expressive: It involves the acts like thanking, congratulating, apologizing, welcoming, etc.
- E. Declarative: It involves the acts of declaration.

iv. Structural Politeness Grade: Like the lexical politeness grade, each structure carries with it certain default politeness value (these values are same as those in lexical politeness). It means that within the same context some kind of structures are considered more polite than certain other kinds of structures. For example, in case of requests an indirect suggestion or a question is considered more polite than an imperative. The rank of these structures are decided on the basis of a survey carried out among the native speakers of Hindi, where they are asked to rank the different kinds of structures in relation to each other on a 5-point scale. The sentences will be annotated for these politeness grades independent

of the lexical items used in the sentence; and it is only at a later stage that the cumulative weighted average of the whole sentence will be computed taking into consideration both the lexical and structural politeness grade.

3. Level 3: POS features – These features are tagged at the level of part of speech and encode the most basic and elaborate information regarding the part of speech in each word.

A couple of examples of the kind of annotation to be done with the features discussed here (summarised, with labels, in Table 1, 2 and 3) are as follows:

- a) [(Rajatji/N_NNP/API !/RD_PUNC Raat/N_NN/NPI ko/PSP/NPI kuchh/QT_QTF/NPI blue/JJ/NPI film/N_NN/NPI typeN_NN/NPI ka/PSP/NPI bhi/RP_RPD/NPI masala/N_NN/NPI dikhao/V_VM/SIL./RD_PUNC)/PRT/SIA/IPM/AFS/DIR/SIM (TRP/N_NN/NPI danadan/RP_INTF/NPI badegi/V_VM/NPI./RD_PUNC)/XE//FUT/SIA/DCM/AFS/AI/NPM/ST1/HT1/ETN/SIT]/RD/OIO
- b) [(Jaan/V_VM/NPI kar/V_VAUX/NPI achha/JJ/SPI laga/V_VM/NPI ki/CC_CCS/NPI tehelak/N_NNP/NPI men/PSP/NPI aise/DM.DMD/NPI bhiRP_RPD/NPI lekh/N_NN/NPI hain/V_VM/NPI jo/PR_PRL/NPI ki/CC_CCS/NPI life/N_NN/NPI vritant/N_NN/NPI hain/V_VM/NPI./RD_PUNC)/PRT/SIA/DCM/AFS/AI/SPM (ab/N_NN/NPI se/PSP/NPI main/PR_PRP/NPI regular/N_NN/NPI is/DM.DMD/NPI patrika/N_NN/NPI ko/PSP/NPI dekhunga/V_VM/NPI./RD_PUNC./RD_PUNC./RD_PUNC./RD_PUNC./RD_PUNC)/EE/FUT/SIA

This annotation is expected to become richer and more complex as we move ahead towards stabilising the scheme.

5. OTHER SCHEMES AND THE REGISTER-BASED SCHEME

Most of the tagsets - like DAMSL (Dialog Act Markup in Several Layers), SWBD-DAMSL (DAMSL applied to the Switchboard data), MRDA (Meeting Recorder Dialog Acts), MALTUS (Multidimensional Abstract Layered Tagset for UtteranceS), DIT++ (Dynamic Interpretation Theory tagset) - to annotate dialog acts (largely of spoken or multimodal data) are based on the assumption that every utterance has some 'communicative functions' and some 'semantic content'. This register-based scheme is largely based on similar assumptions. However the 'communicative functions' are redefined and expanded as 'situational features' in the registeral framework. Within these features the 'communicative functions' (defined in DIT++) are largely covered as the 'communicative purpose' of the dialog acts/ conversation. Besides this the scheme involves annotation of the 'domain' or 'topic' of the conversation which is not covered under other earlier annotation schemes. This scheme also allows for as detailed annotation of 'semantic content' as one wants in the form of 'linguistic features'. Furthermore since the theory proclaims that there is a functional relationship between the communicative functions and the semantic content in one particular register, annotation of this kind may also lead to the prediction of register (or the prediction of the kind of media or place the text belongs to). This kind of functional relationship between 'communicative functions' and 'semantic content' has not been brought out in the previous tagsets; nor are they linked to some particular register or text-types.

6. APPLICATIONS

Any annotated corpus is a very significant resource for NLP in that particular language. Moreover annotation of a corpus with different levels of linguistic information (as well as some extra-linguistic information) is very necessary for the development of any semantics-based tool or application. Till now there have been many isolated attempts at developing annotation schemes at different levels of linguistic representation like POS annotation, syntactic annotation, sense annotation, emotion annotation, etc. But there is no integrated annotation scheme which proposes to annotate the linguistic information at all the levels of representation.

Moreover Hindi is still a comparatively less-explored language from the computational point of view and only a standard POS annotation scheme is available (which is integrated here with this scheme), so annotation scheme given here could be used for the development of several tools, particularly those involving discourse-level processing. It could be especially useful in the development of text classification and text summarisation tools. The very use of the register model implies that certain kinds of texts are associated with prevalence of certain kinds of linguistic features. This feature could be used for text classification purposes.

Hindi does not have a good information retrieval system. So this resource (with required modifications and additions in the annotation scheme) could also be used for development of more powerful information retrieval system, which

would make the Hindi web search engines smarter, more intelligent and much more powerful.

Moreover it could be used for several content analysis jobs like sentiment analysis and opinion mining. One of the examples is the PoRT for which it is being developed now.

Furthermore it could be used for developing other utilities like spell and grammar checker, morphological analyser, etc.

7. ACKNOWLEDGMENTS

I would like to thank my supervisors Dr. Ayesha Kidwai (who introduced me to Linguistics and register analysis) and Dr. Girish Nath Jha (who introduced me to NLP) for their constant encouragement, guidance and support.

8. REFERENCES

- [1] D. Biber and S. Conrad. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press, New York, 2009.
- [2] R. Kumar. Comments in blogs and web portals: a study in politeness. To be presented at 29th South Asian Language Analysis Roundtable (SALA29), 8-10 January, 2011, January 2011.
- [3] E. Miltsakaki, L. Robaldo, A. Lee, and A. Joshi. Sense annotation in the penn discourse treebank. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 275-286. Springer Berlin / Heidelberg, 2008.
- [4] A. Mykowiecka, K. GĄCowińska, and J. Rabięga-Wiśniewska. Domain-related annotation of polish spoken dialogue corpus luna.pl. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC '10)*, Malta, pages 2097-2102. European Language Resources Association, May 2010.
- [5] L. Măărquez, X. Carreras, K. C. Litkowski, and S. Stevenson. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145-159, 2008.
- [6] V. Petukhova and H. Bunt. Towards an integrated scheme for semantic annotation of multimodal dialogue. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC '10)*, Malta, pages 2556-2563. European Language Resources Association, May 2010.
- [7] R. Savy. Pr.a.t.i.d: a coding scheme for pragmatic annotation of dialogues. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC '10)*, Malta, pages 2141-2149. European Language Resources Association, May 2010.
- [8] J. R. Searle. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, Cambridge, 1979.