ISSN (Print) : 0974-6846
**Indian Journal of Science and Technology,** *Vol 8(27), DOI:10.17485/ijst/2015/v8i27/81728, October 2015* ISSN (Online) : 0974-5645

# The TypeCraft Natural Language Database: Annotating and Incorporating Urdu

**Sharmin Muzaffar[1], Pitambar Behera[2*], Girish Nath Jha[2], Lars Hellan[3] and Dorothee Beermann[3]**

[1]Department of Linguistics, Aligarh Muslim University, Aligarh - 202002, Uttar Pradesh, India; sharmin.muzaffar@gmail.com
[2]Center for Linguistics, Jawaharlal Nehru University, New Delhi - 110067, India; pitambarbehera2@gmail.com, girishjha@gmail.com
[3]Norwegian University of Science and Technology, Trondheim, Norway; lars.hellan@ntnu.no, dorothee.beermann@ntnu.no

## Abstract

The authors present one of the important Indo-Aryan languages i.e. Urdu on the TypeCraft platform, which is an online, multilingual, and corpus-based, natural language database and a documentary platform for natural languages. Previously, the platform has already incorporated other Indian languages like Telugu, Bengali, Hindi, and Odia. Recently, the platform has been extended to the annotation and incorporation of Urdu. The TC framework has been designed in such a manner that it can facilitate the linguistic annotation up to the level of semantics to enhance the cross-comparison of structures between languages of different families. The recent version of TC 2.2 has taken the level of annotation up to discourse and pragmatics through a closer integration of text and sentence level annotation. Theoretically speaking, the system is applicable to all languages, but practically it is also very specific with regard to encoding the salient syntactic and semantic features. The paper highlights some of the linguistic issues: Agreement, case, verbs, and mood, labeling features, glossing and technical challenges. The current study focuses on Urdu linguistic annotation taking into consideration the annotated data on the said platform.

**Keywords:** Linguistic Annotation, Natural Language Database, Semantic Argument Structure, South-Asian Languages (SAL), Syntactic Argument Structure, TypeCraft (TC)

## 1. Introduction

India is the home of diverse language families - Indo-Aryan, Dravidian, Austro-Asiatic, Andamanese and the Tibeto-Burman languages[3]. Hindi and Urdu belong to the Indo-Aryan language family and Hindustani is the second most widely spoken language in the world[22]. Standard Urdu is traditionally written in Nastaliq calligraphy style of the Perso-Arabic script, a context-sensitive, cursive and depends solely on Persian and Arabic.

As discussed in[16], TypeCraft is an NLP project maintained by a group of researchers at the Norwegian University of Science and Technology (NTNU), called Research Group in Digital Linguistics, one of the main tenets of which is that methods must be designed in such a manner that they accommodate not only the well-studied languages like English but also the less-resourced and endangered languages. TypeCraft (TC) platform is such a tool. At present, research areas aside from TypeCraft in this group are Grammar Engineering and Language Documentation. In the area of Grammar Engineering, one of the main applications developed by the group is the Norwegian computational grammar NorSource, which applies Head-Driven Phrase Structure Grammar (HPSG) and Minimal Recursion Semantics (MRS) for the advancement of deep-level natural language processing; a related facility is a multilingual valency database, Multi Val[15].

* *Author for correspondence*

Weinreich (1958) in his paper in Word14, 374–9 was the first to use the phrase "convergence area." The term "convergence" has been in use ever since to characterize the phenomenon observed in language contact situations which results in changes at the level of phonology, morphology, syntax, and semantics. It is not clear, however, who first used the specific term "convergence." For a discussion of Trubetzkoy's work, see Velten (1943)

The paper is a demonstration of issues and challenges encountered during the linguistic annotation work undertaken for Modern Indo-Aryan languages: Hindi and Urdu. The issues and challenges witnessed during annotation range from linguistic to technical and from presentation to the description. At present English is used as the meta-language. The structures of the Indian languages are quite interesting, and the investigation keeps an eye open for the possibility that "they can go far beyond the specified parameters of TC Linguistic annotation[16]", and that they can become a challenging task for Natural Language Processing (NLP) tasks like Machine Translation (MT), Content Analysis (CA), Parts of Speech Tagging (POS), Chunking, Parsing, Text To Speech (TTS) systems and so on.

## 1.1 The TypeCraft Platform: An Introductory Background

TypeCraft is a web-based interlinear glossing editor[7] for languages, which utilizes Natural Language (NL) data in the form of sentences and annotate them at nine levels viz. word-level, morphemic, base-level, semantics, Interlinear Glossing (IG), Parts of Speech level (POS), Syntactic Argument Structure (SAS) and semantic argument structure. The native speaker of the respective language or user enters data to the database and adds a linguistic annotation to the written material. It has been designed in such a manner in order to avoid long training periods to work on any language; lesser-described languages or less-resourced languages. So far the Indian languages that have found an entry on the TC platform are Bengali, Telugu, Odia, Hindi and Urdu.

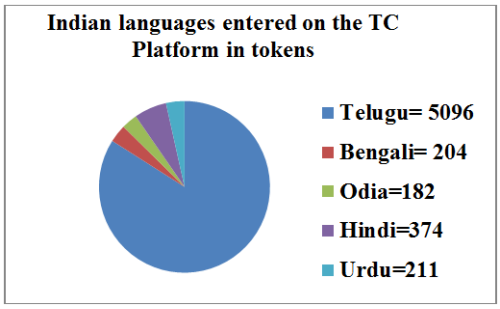


**Figure 1.** Indian languages on the TC platform.

### *1.1.1 Processes of Online Data Entry to the TC Database*

So far as the data entry is concerned, one can either import already structured data or create new texts using the TC Editor to enter text manually and start linguistic annotation. To summarize the levels of linguistic annotation on the TC platform, the following process has to be followed.

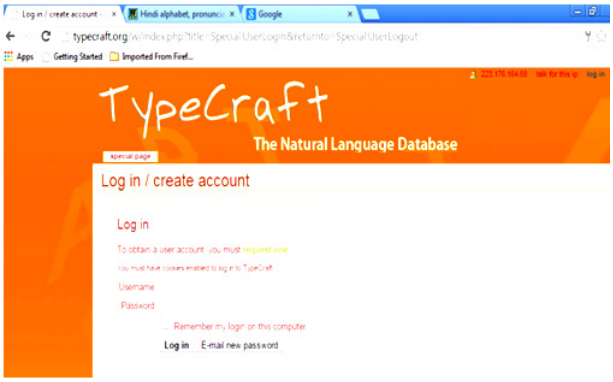


**Figure 2.** TC database log-in page.

### *1.1.2 TC Annotation Process*

- Transcription: To transcribe the text in the concerned language (script of the language), for example, Hindi in Devanagari, Odia in Kalinga and Urdu in Perso-Arabic or in the IPA.
- Sense by sense translation of the text in their own language to English as the meta-language. A second metalanguage can be selected.
- Analysis at the level of word: Analyzing the words in the string.
- Morphemic analysis: Morphemic breaks in the words.
- Base form of words/citation forms.
- Semantic analysis: Meaning of the morphemes in the string.
- Interlinear glossing: With the help of gloss tags list vividly described.
- Parts of speech analysis: POS annotation of the data with the encoding help from the POS tags list.
- Global Tags are the syntactic and semantic descriptions of the string with respect to its syntactic argument structure: Sentence type, argument frame, modality, force and 'evidentiality, sentence aspect' etc. Each one of them contains a list of sub-categories to select the tags from; presented in the drop-down menus.
- Construction description involves the information about the type of construction fed to the TC glossing editor. It does not have a sub-categorical menu and allows the data annotator to give a description name of the respective feature.

## 2. Syntax of Urdu as an Indo Aryan Language

Modern Indo-Aryan consists of Hindi-Urdu, Bengali, Odia, Marathi, Sinhalese and many others[23,51]. South Asia, along with the Balkans, is a paradigm example of the rather rare phenomenon known as sprachbund, or 'linguistic area' or 'convergence area'. Both Vedic Sanskrit and the present Sanskrit are inflecting languages[1]. Since almost all the Indo-Aryan languages have descended from the mother language, Sanskrit, they have inter-assimilated some features in a gradual process in the Indian sub-continent. As a consequence, they share some common features that are unique in nature.

### 2.1 Word Order and Argument Structure

Modern Indo-Aryan languages are verb-final and head-final languages considering the word order. Hindi is a verb final language with grammatical PNG and TAM distinctions. In word order, Hindi and Urdu display some features of a verb-medial language. There is a clearly observable distinction between the order of constituents in English and Hindi. For example, unlike in English, the spatio-temporal circumstances of an event are mentioned before the arguments involved in the event, e.g., word order is relatively free, since in most cases postpositions mark quite explicitly the relationships of noun phrases with their constituents of the sentence. As a result, for the purposes of thematization and contrastive focus, constituents can be moved around freely within the clause[18]. This is similarly true to the case of Urdu language.

With respect to the transitivity/intransitivity of the verb, three types of structural patterns are feasible: NP+VP (when the verb is intransitive), NP+NP+VP (when the verb is transitive), and NP+NP+NP+VP (when the verb is ditransitive). Hence, the verb can have three arguments structurally. The adjuncts are generally placed before the verbs in most of the Indo-Aryan languages except the cases of interrogative and negative sentences where the right-ward displacement takes place. "All South Asian languages except Khasi are left-branching". English is right-branching. In South Asian Languages (SALs) the auxiliary verb follows the main[24]. Thus: Verb = Main Verb + Auxiliary.

### 2.2 Case

Trask has said that, 'one of the forms which a noun or pronoun may assume in order to represent its grammatical and semantic relation to the rest of the sentence'. This section aims to explore the case marking system in Urdu. In most of the Indo-Aryan languages, case is realized in the form of postpositions; especially when they take nouns grammatically from phrases. Therefore, they are called postpositional phrases. These sorts of phrases consist of noun phrase followed by a postposition.

From the perspective of case, languages are of two types: Nominative-accusative and ergative-absolutive. Languages may also differ in terms of case marking on the subject and object. For instance, the subject may be case marked by the nominative case in some languages or it may be case marked by a case marker which is ergative. The former is known as 'nominative-accusative' type (e.g., French, German, and English) and the latter as 'ergative-absolutive languages' (for example, Georgian, Hindi–Urdu, and Punjabi).

In an ergative language, the subject of the intransitive verb and the object of a sentence with a transitive verb are case marked identically and the subject is marked with the ergativity. The subject with an ergative case usually exhibits the syntactic properties of a subject$_{22}$- when there is ergativity in a sentence, sometimes the

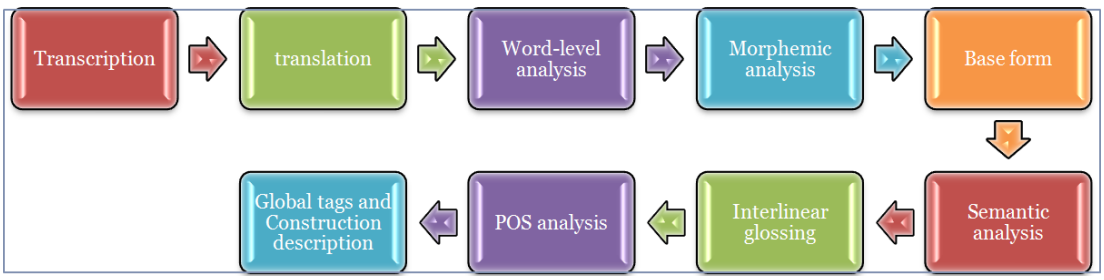


**Figure 3.** TC annotation process.

agreement triggered is object-verb agreement. In Hindi and Urdu both the nominative and ergative construction of case is prevalent. Further, in Urdu, there are ergative-nominative and ergative-dative alternations[10]. In addition, in the perfective aspect, the ergative sentences can be expressed with nominative cases. To put forth in other words, the same syntactic construction can have each two ways of representation. First phase consists of ergative-nominative and ergative-dative constructions while the second consists of nominative-ergative alternations.

### 2.2.1 Ergative-Nominative and Ergative-Dative Constructions

mɛne roʈi kʰɑi

I-1.SG.MASC.ERG CHAPATTI-1.SG.FEM.NOM eat-PST.PFV

"I have taken/eaten chapatti".

mɛne ʊsko roʈi kʰɪl-ɑi

I-1.SG.MASC.ERG him-DAT CHAPATTI-1.SG.FEM.NOM feed-PST.PFV

"I fed him/her chapatti".

### 2.2.2 Nominative-Ergative Alternations

The sentence mentioned below in the ergative case can be also represented with a sentence having the nominative case.

ʊo log kʰɑnɑ kʰɑe

those persons-3.PL.NOM. food eat-PL.PST.PFV

"They ate food".

ʊn logõne kʰɑnɑ kʰɑjɑ

those persons-3.PL.ERG. food-3.SG.MASC eat-PST.PFV

"They ate food".

One of the interesting things that is noticeable here is that for both the nominative and ergative cases in Urdu, the same English counterpart has been used.

In Case markers may be preceded by the definite or plural marker. The accusative case is used only when the direct object is specific[24]. According to[12], 'the ergative case in Hindi-Urdu is a structural case and the dative and other non-nominative cases are inherent (lexical) cases'. That is, the latter are 'theta-related' and selected by the predicate depending on the nature of the predicate. The position of Davisson is in contrast to Hook, who argues that ergative case in Hindi-Urdu is an inherent case[2].

### 2.2.3 Case Marking

Case marking is typologically postpositional in nature in Indo-Aryan languages. The subjects of the finite intransitive clauses receive nominative case which is unmarked generally. The nominative case seems to be licensed by finite tense in many Indo-Aryan languages. A nominative subject cannot appear in a nonfinite clause in Hindi. That is why; marking genitive case is an option that is generally available along with the dative, accusative, and the non-nominative constructions, in some special cases.

#### 2.2.3.1 Oblique Case

The oblique is a grammatical or structural case having no semantic content, but the agreement is usually triggered when nouns and pronouns are preceded by the objects of the postpositions. Adjectives are marked for the oblique case in agreement with the nouns which they modify or qualify. For instance, in the following example, /əcc$^{h}$e/ 'adjective' depends upon the number of the nominal it marks i.e., /kele/ 'banana'.

Adjectives are marked for the oblique case in agreement with the nouns which they modify or qualify.

ʊsne əcc$^{h}$e kele lɑye

he-3.SG. MASC.NOM. good bananas bring-PAST.PFV

"He brought good bananas".

#### 2.2.3.2 Quasi-Ergative and Ergative Agreement

In some MIA languages the verb agrees with both the object and the subject at the same point of time which is one of the major issues as to how to assign case and where to mark the agreement. Thus it can be stated that these languages sometimes trigger the double agreement structure and a cross-universal merger of nominative-ergative case marking. Therefore, it can be stated that it is the case assigned on the subject of a finite transitive clause in the Western belt of Indo-Aryan languages. Hindi and Urdu are ergative-absolutive languages, while the eastern group of Indo-Aryan languages is nominative-accusative languages. The ergative case is marked on the subject by the case-clitic /ne/ in the perfective aspect while in other circumstances it is dative-subject or the nominative case which prevails the most[24].

usne pɑnI pine kI koʃiʃ kI

he. 3 SG.MASC.ERG water.ACC to drink try.3SG. NOM.Vcon

"He tried to drink water".

#### *2.2.3.3 Non-Nominative Subjects or Dative Subject Constructions*

The Indo-Aryan language exhibits a wide-range of constructions where the subject receives a non-nominative case apart from the ergative constructions[21]. This argument has been called a subject because it meets a subset of subject hood tests. In the example the subject is *he,* which is in the objective form of the pronoun, but the object, that is, *thirst* seems to possess the role of the agent. Consequently, the final triggered case is of the non-nominative construction type.

NP + NP - E x p r i e n c e d E m o t i o n — Adverb Phrase: Manner

ek ɖʱn ʊs-ko bɔhɔtʰ ʐor se pjɑs ləgi

one day he-DAT very thirst-3.SG.FEM.PAST.PFV

"One summer day, he felt very thirsty".

## 2.3 Verbs

In MIA languages, Verbs are inflected for PNG and TAM features with both subject and object in Urdu. Verbs take one derivational class of affixes: The causal affixes for the first and the second causal. Syntactically, verbs determine the number and function of noun phrase arguments in a sentence. With regard to semantics, "they express states, processes, and actions[3]". The basic verbs as well as causatives behave identically with respect to aspect, mood, tense and agreement features.

Inflected Forms of Urdu verbs:

Morphologically, the verbs have the following forms:

**Table 1.** Inflected forms of Hindi verbs (as adapted from Kachru[17] )

| Verb class | Verb form examples |
|---|---|
| Root | ɖɛkʰ 'see, look' |
| Imperfect participle | ɖɛkʰtɑ |
| Perfect participle | ɖɛkʰɑ |
| Causative | ɖikʰɑ (first causal) dikʰwɑnɑ (second causal) |
| Infinitive | ɖɛkʰnɑ |

The causative derivation increases the valency of the verb, i.e., it adds one more argument to the argument structure of the verb. There are lexical causative verbs in Hindi such as /ɖikʰɑ/ (first causal) 'to show', /ɖikʰwɑnɑ/ (second causal) 'to cause someone to show'.

### *2.3.1 Serial Verbs*

Serial verbs are the verbs which describe the serial occurrences of the process or action, where the lexical level is one main verb and other auxiliary verbs occur in a serial manner. The nonfinite past participle is formed by adding the suffix /-kər/ or /ke/ to the verb stem which is also the conjunctive marker. A series of past participles, one followed by another also occurs in Odia. Such formations are referred to as serial verb constructions. In Urdu, the list of verbs that can occur as auxiliary verb is also greater in number than many other languages like English. Given a serial verb construction as for example in the following construction:

ʊəh məhinõ̃ se pəʈʰ-ʈɑ cəl-ɑ ɑ rəh-ɑ hɛ:

he-3M.SG.NOM months-PPOST read-IMPF.M.SG walk-IMPF. come-VB live-PFT.MSG be-PRS.SG

'He has been reading for months'.

We find that a sentence can have several verbal elements put together and still denote a single verbal process i.e. it still has one main verb whereas the others function as auxiliary denoting other features like tense, aspect and mood. In the sentence above we find that the main verb /pəʈʰ/ 'to read' is followed by four other words with verbal intent and act as auxiliaries, having different tense, aspect, mood and agreement markings. Also it is worthy to note that these words separately contain different aspects and tense i.e. the TAM and agreement is spread over all these elements.

### *2.3.2 Compound Verbs*

Compound verbs are a sequence of two verbs AB ((main verb A) plus *vector B*). It imparts completion (Butt, 1995) or the attitude/feeling of the speaker towards the event. The general structure of compound verb as stated above is V1+V2. In this structure, it is the V2 that takes the TAM and the agreement markings while V1 remains in the root form. V2 has also been called with several other terms such

as 'vector verb' and 'light verb'. V2 is a small set of words that has been listed variously in different descriptions of Hindi verb structure. There is no unanimous agreement among linguists on how many light verbs in Hindi exist. Hook reports that there are studies that list these verbs starting from having a total number of 8 to 61. The following verbs playing as V2 are commonly regarded as auxiliary verbs that represent aspectual or modal values: /rəh/, /sək/, /cʊk/, /pɑ/, /ləg/, /ɖe/, /ɟɑ/ and /kər/ and so on.

ʊo pɑni pɪ-ke ʊɽ gəjɑ

he-3.SG.MASC.NOM water-3.SG. drink-PFV fly-SG. SG.PAST

PN NcommV1 V2

"He flew away after drinking water".

Besides these single words, there are verb sequences also that may occur as part of the compound verb, making in totality of three word compound verb constructions. It is notable here that since such sequences, we have adjectives like/kʰəɽɑ/ or adverb like/bɑhər/ as first component of such sequence while the following word is the normal V2. This is a property typical to the conjunct verbs.

### *2.3.3 Conjunct Verbs*

Conjunct verbs are formed of a nominal or an adjective followed by a verb. These two together semantically denote an action or a process or a state. The verbs that take part in the conjunct verb construction consist of a small set. The members of this set are /ho/ 'be, become', /kər/ 'do', /de/ 'give', /ɑ/ 'come', and /ləg/ 'apply'. The process is very productive; any noun or adjective can be used in this construction to yield a corresponding verb[24]. Morphologically productive usage of noun/adjective plus a light verb as predicate. It is the light verb that 'carries the tense, aspect and agreement markers'. For instance, / telephone kərnɑ/ 'to telephone' in Hindi-Urdu:

mɛ ne rɑm kɪ mədɖəd̪ki

I-1.SG.M.erg Ram gen help-FEM.SG did. FEM.SG

'I helped Ram.' (Masica, 1993)

ɪs ʊəqt̪ʊse ek t̪ərqib sʊɟʰi

this time he one idea-3SG.FEM. come-CONV

"Suddenly he-DAT came up with an idea".

In the above-mentioned example /t̪ərqIb sujʰi/ 'came idea' is the conjunct verb.

In[21] all the Dravidian languages and in many Indo-Aryan languages (Assamese, Bangla, and Odia), the patient of the conjunct verb takes a dative case marker verb. Enriched usage of noun or adjective with a light verb as predicate. It is the light verb that carries the tense, aspect and agreement markers. For instance, /telephone kərnɑ/ 'to telephone' in Hindi-Urdu is the conjunct verb.

### *2.3.4 Conjunctive Participle*

ʊsne ek kəŋkəɽ kɑ ʈʊkɽɑ məʈke mɛ̃ lɑ-kər ɖɑlɑ

he-3.SG.MASC.NOM a piece of pebble in brought-CONJP drop-3.SG.MASC.PAST.PFV

"He dropped a pebble into the pitcher".

In Hindi-Urdu, the case of conjunctive participle is slightly different from that of Odia with respect to the presence or absence of the participle markers. Here in the language for a verb to be a conjunctive participle candidate, it has to have the markers like /kər/ and /ke/. The finite verb seems to carry the inflections for all the TAM features and person-number-gender markers.

## 2.4 Agreement

There are two types of agreement in Hindi: Subject-verb agreement and object-verb agreement:

The former is a direct agreement as the verb directly agrees with the subject based on the PNG of the subject in question and TAM of the sentence. When there is a direct agreement, usually the subject who is the agent of the action gets marked for its agreement; there occurs a nominative case. In the latter, when the subject is in the dative-subject case or in the ergative case, the verb agrees with the object of the sentence and accordingly is inflected for the PNG and TAM features.

Hindi phrases and sentences show two types of agreement patterns; 1. Modifier-head agreement and 2. Noun-verb agreement[17]. "Modifiers, including determiners, agree with their head noun in gender, number" and case, and finite verbs agree with some noun in the sentence in PNG.

Kisi ek ləʈke-ne gɑnɑ ʃuru kiyɑ.

some.OBL oneboy.M.SG.OBL AG singing.M begin do.PERF.M.SG

"Any one of the boys has started singing".

## 2.5 Mood

The moods that are marked morphologically on the lexical verb itself are imperative, optative and contingent. Others are formed by the concatenation of infinitival or participial forms of verbs and aspect-tense auxiliaries.

The morphology of aspect-tense-mood is complex.

Language not only encodes information about entities, relations, temporality, locations, etc., it also signals expressive and social information. The part of grammar that encodes the social and the expressive is the mood system. The distinction between indicative and imperative, for example, signals the different social values of statements and commands, respectively. Similarly, the distinction between indicative and presumptive encodes what the speaker's perspective is about the situation, i.e., whether he/she asserts it as 'real' or 'to be presumed to be real' on the basis of relevant evidence available to him/her. In Hindi, within the verb phrase, a six-way mood distinction is made: indicative, imperative, optative, presumptive, contingent and past contingent or counter-factual. These are expressed by the following forms (all the forms, except the example for imperative and optative, which are not marked for gender, are in masculine singular):

Activity-declarative-dubitive-serial Verb Construction-adverb Phrase: time--cognition-S+NP+S

usne socɑ kɪ əb pɑni kɛse pɪɑ ɟɑe

he-3.SG.MASC.ERG think-PAST.PFV now how to drink-IMPER water

"He thought how to drink water now".

The imperative mood is the root form of the verb without being marked for PNG. The optative mood is presented by an inflection on the verb which additionally indicates the person and number of the subject. The other moods are indicated by a concatenation of the participial form of the verb with an inflected form of the auxiliary. All the forms cited above in, except for the imperative and the optative, are third person singular, and additionally, all forms involving a participle are masculine; the participial forms consist of the present or past participle, or the progressive form of the verb. In the indicative, interrogative and negative, the verb root or aspectual form is followed by either the present or the past auxiliary. In presumptive, contingent and past contingent, the verb root or aspectual form is followed by the auxiliaries' / hogɑ/, /ho/ and /hoʈɑ/, respectively. All these forms are discussed in some detail in the following sub-sections.

## 3. Technical Challenges

LTs (Language Technologies) are the technologies that facilitate the communication of complex information; so they are considered to be Human Language Technology. Language can be of two forms: Written and spoken. While there is consistency in the former, the latter exhibits no such thing. When there is a spoken corpora selected for the testing of the newly-created model, it may prove to be not so advantageous for the written text corpora; thereby creating an NLP challenge.

There has been always a challenge for the linguistic theories for the description and accommodation of all the existing languages. While attempting to incorporate all the languages of the world with theories, we miss out some every time. Based on the existing theories, while envisioning for a platform, it becomes a primordial challenge for computational linguistics to accommodate all the features of all the existing languages. Consequently, there has been always a discrepancy in integrating the theories and analyzing from different levels of linguistic structures and model comprehensive and accommodative systems of representation.

One of the ultimate objectives of the LTs is the translation of human languages automatically. Whereas, multilingualism in a society is well-advantageous for languages, the unprecedentedly growing multilingualism on the web creates significant challenges for the LT. Apart from this; the modeling of language is trans-disciplinary in nature as it encapsulates diverse areas of human knowledge[26]: Computer science, computational and theoretical linguistics, mathematics, electrical engineering and psychology.

Attempting to address the challenges will require 'massive scaling-up' in terms of data size and examine our hypothesis in linguistics; besides developing new computational methods. As stated by[9], the challenges are the data acquisition, data mining, and the complexity challenge. The possible solutions could be 'uniform standards', 'standards-compliant tools', 'coordination and collaboration' among stakeholders and linguists, 'data sharing', and 'computational methods'.

## 4. TC Linguistic-Technical Issues

SAL as a linguistic area is already averred by Emeneau, which exhibits unity in diversity. In other words, although there is the existence of five major groups of languages: The Indo-Aryan, Dravidian, Austro-Asiatic, the Andamanese and Tibeto-Burman; there are some unitary features. MIA languages are inflecting in nature where the verb inflects for TAM and PNG features. This inflection along

with other distinguishing features leads to a significant challenge for the field of NLP.

Firstly, for the non-nominative subject construction, the subject is the experiencer of the action. So, this aspect is one of the salient features in SAL and MIA languages. If one marks the subject as the NOM, there is a loss of semantic feature that the subject should possess. However, the TC platform allows one to specify the subject as both experiential and dative, through the composite tag 'EXP.DAT (a dot separating the independently defined tags). Annotating experience subjects is thus not a challenge for TC.

In a related vein, dative-subject construction is one of the unique features of the MIA languages[17]. When the logical subject, unlike the grammatical subject, is a possessor or an experiencer of the action being performed, the subject is marked for the dative case. So, there should be specified tags provided to the annotators, or else there is a misconception between the dative case where one of the objects being the benefactor of the action and in the case of logical subject being the experiencer of the action. Also for this situation TC provides a solution, in that the composite gloss tags 'SBJ.DAT' and 'OBJ.DAT' are available, in the same way as above.

## 5. Conclusion

The issues that crop out of the study relate to the uniquely interesting structures of the MIA languages like the dative-subject construction, conjunctive participle, non-nominative subject construction, are in fact catered for by TC, however a more comprehensive discussion of the use of complex tags is needed. More attention has to be paid to issues involving semantic classes such as psychological verbs of emotion, physical state, and mental perception, and existential, which are instrumental in the construction of a variety of sentence types, including nominative, dative, non-nominative, dative-subject, ergative case and so on.

Another issue to be noted here relates to the serial verbs. Urdu abundantly use complex predicates formed from a variety of sources, including Sanskrit, Persian, and Arabic. In the case of a serial verb of any sub-type, the main verb becomes dormant in some cases and sentence aspect/tense affixes and verbal compounds define the interpretation as state, achievement, activity or accomplishment; in combination with some core verb meaning. Further, the explicators, the auxiliaries, copular verbs, vector verbs are inflected and bear the markers for mood, tense, aspect and some other grammatical features. A common annotation schema to cover this wide area of issues mentioned will be a goal of consolidation and future research.

In the foregoing it can be averred that for further research and development work collaboratively, the data can be imported and exported between the two platforms (JNU and NTNU) so as to facilitate the process of linguistic annotation. On one hand, the TC platform will benefit from the incorporation of the Indian languages while on the other, the ILCI will get the additional levels of linguistic annotation.

## 6. References


1. Abbi A. Semantic grammar of Hindi: A study in reduplication. New Delhi: Bahri Publications; 1980.
2. Abbi A. The conjunctive participle in Hindi-Urdu. International Journal of Dravidian Linguistics. 1984; (13):252–63.
3. Abbi A. A manual of linguistic field work and structures of Indian languages. Lincom Europa; 2001.
4. Agnihotri RK. Hindi: An essential grammar. Routledge; 2013.
5. Beermann D, Mihaylov P. TypeCraft–glossing and data basing for linguists of linguistics. 2008; 11.
6. Beermann D, Hellan L. TypeCraft: A natural language database. Legon-Trondheim Linguistics Project Meeting in Accra; 2006 Jan.
7. Beermann D, Prange A. Glossing language online. Proceedings of the Texas Linguistics Society X Conference Computational Linguistics for Less-Studied Languages. Stanford: CSLI Publications; 2006.
8. Hellan L, Beermann D, Bruland T, Dakubu MEK, Marimon M. MultiVal–Towards a multilingual valence lexicon. LREC; 2004.
9. Bender EM, Good J. A grand challenge for linguistics: Scaling up and integrating models. White paper contributed to NSF's SBE, 2020; 2010. 1-1. Available from: http://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Bender_Emily_81.pdf
10. Butt M, King TH. The status of case. Clause structure in South Asian languages. Springer Netherlands; 2004. p. 153–98.
11. Das PK. Grammatical agreement in Hindi-Urdu and its major varieties. Lincom Europa; 2014.
12. Davison A. Lexical anaphors and pronouns in Hindi / Urdu. Lexical Pronouns and Anaphors in Selected South Asian Languages: A Principled Typology; 2000. p. 397–470.
13. Hardie A. The computational analysis of morpho-syntactic categories in Urdu [PhD thesis]. University of Lancaster; 2003.
14. Humayoun M. Urdu morphology, orthography and lexicon extraction [Master thesis]. Chalmers University of Technology; 2006.

15. Hellan L, Beermann D, Bruland T. Towards a multilingual valence repository for less resourced languages. Proceedings from the 4th Language Technology Conference; Poznan. 2013.
16. Jha GN, Beermann D, Hellan L, Singh S, Behera P, Banerjee E. Indian languages on the TypeCraft platform: The case of Hindi and Odia. Proceedings of the LREC; Rekyavik. 2014.
17. Kachru Y. Hindi. John Benjamins Publishing; 2006.
18. Kachru B, Kachru Y, Shikaripur SN. Language in South Asia. Cambridge University Press; 2008.
19. Kumar R, Kaushik S, Nainwani P, Banerjee E, Hadke S, Jha GN. Using the ilci annotation tool for pos annotation: A case of Hindi. 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012); New Delhi, India. 2012 March.
20. Masica CP. Defining a linguistic area: South Asia. Orient Blackswan; 2005.
21. Muzaffar S, Behera P. An error analysis of the Urdu verb markers: A comparative study on Google and Bing machine translation platforms. Aligarh Journal of Linguistics. 2014; 4(1-2):199–208.
22. Masica CP. The Indo-Aryan Languages. Cambridge University Press; 1993.
23. Rahman T. Language policy and localization in Pakistan: Proposal for a paradigmatic shift. In SCALLA Conference on Computational Linguistics; 2004 Jan.
24. Subbarao KV. Typological characteristics of South Asian languages. Language in South Asia; 2008. p. 49–78.
25. Subbarao KV. South Asian languages: A syntactic typology. Cambridge University Press; 2012.
26. Schmidt RL. Urdu: An essential grammar. Routledge. In: Uszkoreit H, editors. Language technology a first overview. German Research Center for Artificial Intelligence; 2005. p. 1–4.
27. Uszkoreit H. Language technology a first overview. German Research Center for Artificial Intelligence; 2000. p. 1–4.