The Generative approach to the computation of Basic verbal-strings in Hindi *

Pradeep Kumar Das**

Abstract:

The paper examines the possibility of developing a computational parser for verb morphology in Hindi that would generate correct verbal stems for different kinds of tense and aspects. Such computational approach is needed not only for Hindi but for many other inflectional languages such as Latin, French, Spanish, Punjabi, Marathi, Korean etc where students are burdened with, or rather forced to memorize a huge data-set of verb conjugation. I personally believe that this kills the beauty of learning an inflectional language. The burden can be at least minimized by this generative approach.

Hindi is used in the paper just for the corpora, however, the model is supposed to be an open ended and has the power provide the correct computation for other languages too. The conceptual framework remains the same for every inflectional language; however, the degree of inflection decides the variables that one has to parameterize in order to make the model work for a particular language.

The claim might look a little big at the moment, but I personally have already used it for several Indian languages (about fifteen) during my research work for PhD. The degree of variation with regard to the inflectional markers in my work formed a very neat and understandable continuum and the generative approach helped a lot to understand the huge data-set of the verbal strings that I needed to compare and contrast in order to find out the systems of agreement in the major varieties of Hindi-Urdu.

Exp: The following ONE generative computational rule can produce all required verbal strings for Present Indefinite Tense in Hindi:

Aspect Sub/ObjProp tense
Verb root + **t** + **PNG** **of Sub/ obj** + **hona (**=PNG of Sub/ obj)

---

0. Introduction
1. The basics of computational methods
2. The system of verb agreement in Hindi
3. The notion of time, tense and aspect
4. The computational parser for basic verb-stems in Hindi

* This work (research paper) is supported by Hankuk University of Foreign Studies Research fund in the academic year 2008-2009

* The paper was presented in the conference "**Global Association of Indo-ASEAN Studies**" held at Pusan University of Foreign Studies, Pusan, South Korea, 14-11-2008.

** The author is an Associate Professor of Linguistics at the Department of Linguistics, University of Delhi, Delhi, India

0. Introduction:

The paper is divided into four parts. The first part lays out the foundational and basic concepts of computational methods. The second part explains the types of verb agreement in Hindi and other inflectional languages (Indian languages). The linguistic jargons that are used in the paper to establish the agreement system are not very difficult to understand and do not call for any expertise in linguistics. The third part tries to establish the conceptual sketch of the notion of *time*, *tense* and *aspect* in general. Since this background information is attached as the backbone of the parser in the system, it has all its simplicity and self-explanatory details and should not pose any problem for the general readers to understand it. The fourth part of the paper provides details for the application of the parser which demonstrates on paper how the parser will help the computer to compute all desired values of tense and aspect of the verbal stems in Hindi and other inflectional languages. A word of caution must be put forward before I move into the explanatory aspect of the parser. The machine based translation is not a simple task in the field of computational linguistics. Something that looks so simple and easy for us and we feel kind of irritated when we do not see any pre-conceived related analysis of the part of linguistic problems. The caution that I wanted to refer to is this that a linguistic problem might look very simple to us, however, it really takes time for people working in the field of computational linguistics to map very simple facts by tagging the variables in the parser for the so called simple linguistic problem. It might be useful and encouraging if we show bit of endurance and agree for a while to what the computational linguists are doing, and do not try to impose our frame of mind and irritated anxieties of not being able to find their preconceived explanation of any linguistic phenomenon. I also need the above mentioned endurance and tolerance for my explanation of the notion of 'tense and aspect' system that is found in the languages. I am fully aware that I have tried to adopt the Anglo-centric approach to the concept and have probably ignored many other well-established terms such as 'Imperatives, Habituals and other linguistic jargons' that are so much in the mind of the people who deal with Indic languages. I, however, would like to repeat it again that let the computational linguists use the model that is explained here, and if the success of the parser is achieved to the expected level, we can add and modify the terminology very easily. It is a new way to look at the problems and there has been no such prior attempt to develop this kind of PARSER in the literature for the verb morphology of Hindi and other inflectional Indian languages.

1. **The basics of computational methods**

The morphological parsers that are developed for various kinds of computations suffer the drawbacks of not having sufficient resources. Though, it is also true that parsing morphologically rich languages and with relatively free word-order (e.g. Hindi and some other Indian languages) has always posed challenges for the computational linguists. However, what is known as the 'dependency parser'[1] has proven very useful for parsing data in morphologically rich languages. The dependency parser is generally divided into two types, 'grammar-driven' and 'data-driven' parsers (Carroll 2000). The grammar-driven dependency parser works by eliminating the tags which do not meet the given constraints. In other words, the grammar-driven parser sees parsing as a constraint-satisfaction process. The data-driven dependency parser, on the other hand, uses the corpus of a language to induce a probabilistic model for selecting the correct form or string. The present parser is modeled on 'data-driven' dependency logic. The philosophical base and the linguistic technicalities of the parser would help the computer to identify correct strings of the verbal stem and label the appropriate tense and aspect markers for the intended expressions. However, it is important that the team responsible to develop this parser must consist of mathematicians, computer software engineers and language experts. There has been a growing interest in the field of computational linguistics to attain various kinds of parsers that would do the simple parsing for the linguistic corpus and make several kinds of language analysis quite easy and simple. In the long run, such parsers should enable us to translate the given English text in Hindi or in any other inflectional language with minimum or no errors of grammar. I personally believe that the computational linguists must be appreciated for their hard work that they put in to make our life comfortable. They have not been able to get the due acknowledgement in many occasions. Hence, we must be grateful to all those who have worked behind the screen and made this possible to facilitate us with the great developments in the field of computers, especially in natural language processing.

It is said that there are at least five stages of technological conversion that must take place within a fraction of second when we press a key-button of the keyboard and the monitor shows up the typed latter on the screen. These stages are as follows:

[1] See Bharati et al. (1993, 1995) for detail account of the 'dependency parser'.

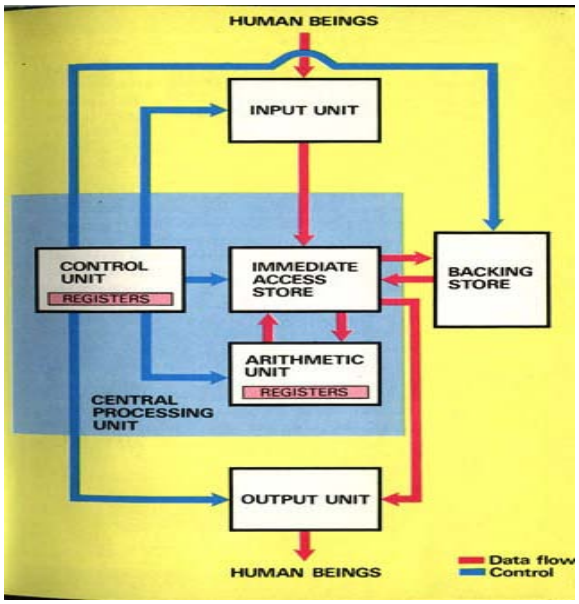
1.Keyboard to Mother-board (magnetic field of the key-board changes the power energy to electronic energy)

2. Mother-board to Process (non-binary to binary process)

3. Processor to Hard Disk (mechanical signal to optical signal)

4. HD to RAM (static signal to digital and free-floating data)

5. Optics to Visual (Non visual to visual graphic)



http://pointlessmuseum.com/computer/007.html

This could not have been possible if the computational linguists would not have given their sweat and blood to develop different kinds of computational tools and mechanisms. It is for this reason that I feel indebted to all those who have come forward and helped to conceive the notion of this parser, called 'morphological parser for the verbal stems in Hindi and other inflectional Indian languages'.

2. **The Agreement System:**

In this second part of the paper, I would lay out the basic agreement system available for Hindi and other inflectional languages. Look at the diagram and the details given below. You need to have patience in order to make it convincing to you. Despite all my efforts, it might look a little bit complex in its outfit, however, it is going to help the computer to match or compute the correct values of complicated grammatical properties. So, look at the diagram carefully:

Constituent $\mathbf{Y}_{(verb)}$ agrees with constituent $\mathbf{X}_{(NP)}$ (in a category **Z**) iff the following conditions hold true:

a. There is a syntactic relationship between **X** and **Y**.

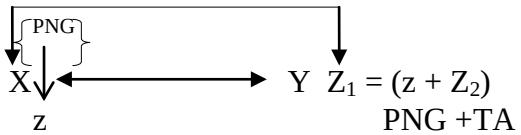
b. **X** belongs to a subcategory **z** (the nominal inflections) of a grammatical category **Z** (the verbal inflection), and **X**'s belonging to **z** is independent of the presence or nature of **Y**.

c. There is a mutual sharing of the grammatical features between the nominal and verbal inflections in the clause.

d. **z** is expressed on **Y** (as a verbal inflections i.e. **Z**) and forms a constituent with it (i.e. VP = verb phrase). However, in case of no suitable X is found, the Y selects its own z and makes up the VP. The value of 'z' in this case is fixed and it is 3PMS in Hindi. Das, P.K. (2006)

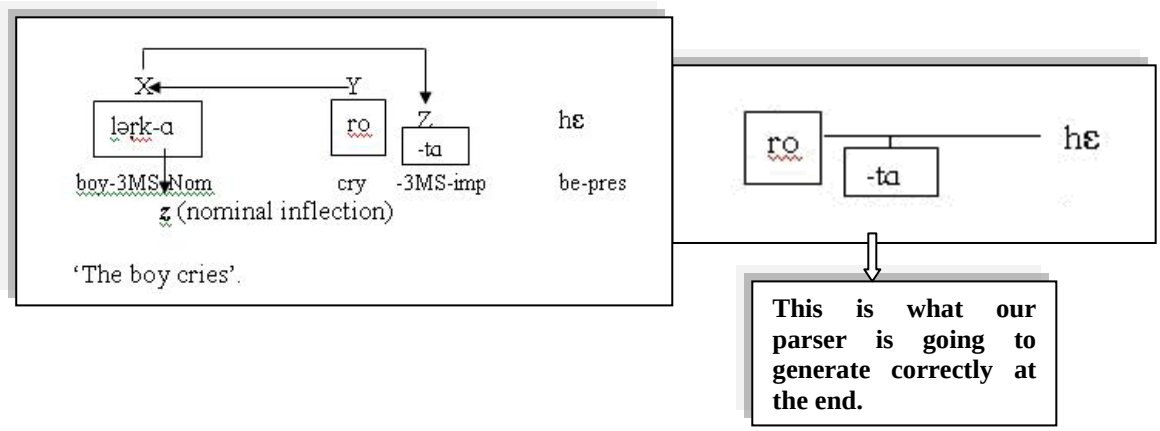
Now, the above definition will give us the following schemata for various kinds of agreement systems available in Hindi and other inflectional Indian languages:

1. Subject –verb Agreement:



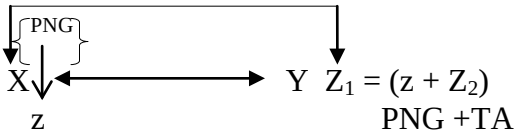
Feature-values:
X= subject
Y= verb
Z_1= inflectional marker
z= Nominal inflection
Z_2= Tense and aspect

Corpus example:



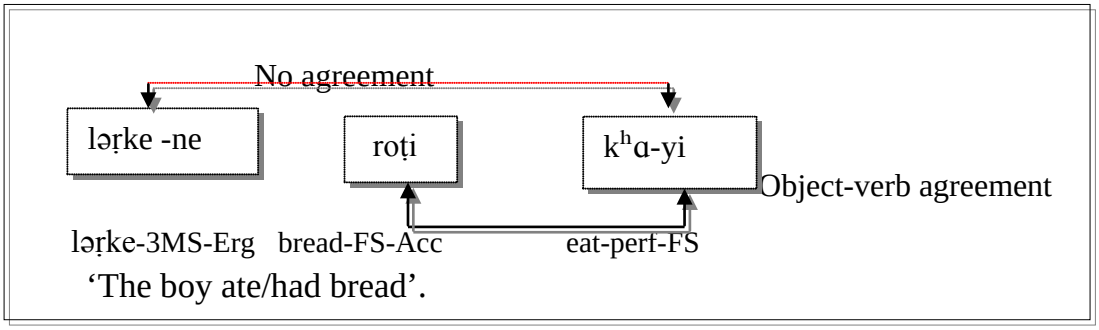
The second sub-system or scheme that the aforementioned definition can bring to meet the requirement of the agreement of the language can be shown in the following diagrammatic form:

2. Object-verb Agreement:

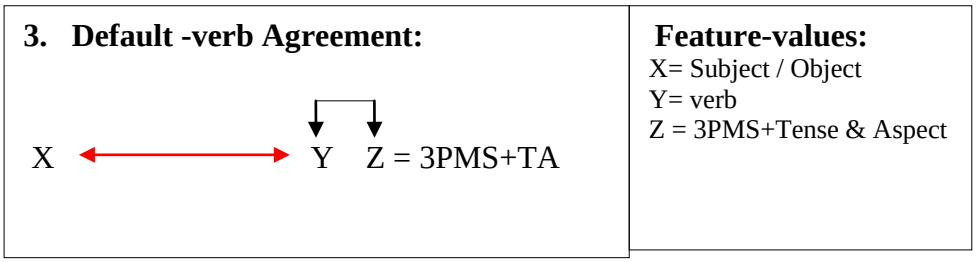


Feature-values:
X= Object
Y= verb
Z_1= inflectional marker
z= Nominal inflection
Z_2= Tense and aspect

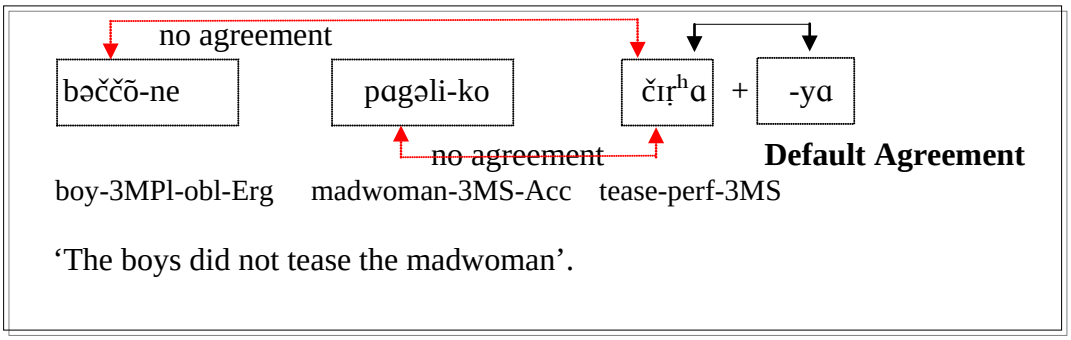
Corpus example:



The third schema that is derived from the aforementioned definition of verb agreement is actually specific and parameterized for the purpose of meeting the verb agreement of Hindi in particular. The third type of agreement is called the 'default agreement' in the literature. The feature value of the default agreement in Hindi is fixed and it is always third person masculine singular i.e. 3PMS. It is obvious that in this case the verb does not find any eligible noun phrase to agree with and thus it acquired this default form of agreement. In the technical terms, there is no relationship between the X and Y and in such case, the Y (VP) has its own Z (verbal inflection) to turn the verb into finite one.
Look at the diagram given below:



Corpus example:



The basic layout of the verb agreement in Hindi and other inflectional languages serves as one of the backbones for the parser. The complex feature-values such as

person, number and gender of nominals in the sentence and the mapping of these features on the verb phrase is mediated by the help that the parser extends to the whole process of computation. However, before we move on and talk about the actual execution of the parser, it is important to explain other inflections i.e. verb inflection which will have to combined with nominal inflections to complete the verb phrase. The verb inflection is made out of the tense and aspect markers. So, we must explain the notion of tense and aspect in the following section:

3. **The notion of Tense and Aspect:**

It is difficult to define the notion of tense in absolute term. This is so because it requires us to know the concept of time that is technically needed to understand the term 'tense'. It is assumed that 'time' is a continuous flow of events. An event is an incident that is actually an action and is mostly denoted by the verb. So, if we want to understand what the term 'tense' means, we must propose a working definition of tense by saying that 'tense is an effort to locate or identify an action in the time scale'.

However, it appears that it is not so simple to categorize the notion of time beyond what we said earlier i.e. '*it is a continuous flow of events'*. This is partially so because no one can claim to know the starting and the end or any such referential point with regard to the notion of time. Yet, we find ourselves so confident in everyday life in talking about so many things that might refer to various points of references in time-scale.

It appears that this facility has come to us from grammar, where time is seen as a construct which holds important knowledge of the point of reference of an action. It is also important to note that if we do not have a shared knowledge of time, it would be literally impossible to have any communication in any given language.

So, there must be some way to handle this technical aspect of the notion of time that seems to work as the underlying form to let the notion of tense emerge as one of the basic components of universal grammar. If we think carefully, we would come to this consensus that there must be the following three factors that can help in establishing the overall, though individual, knowledge of time:

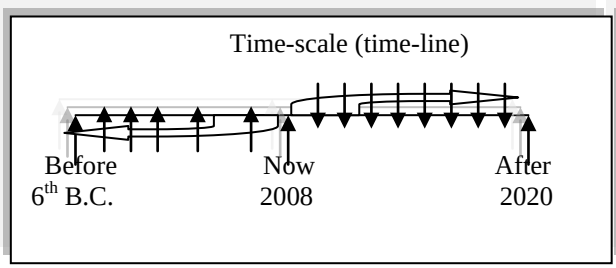
a. Knowledge of time through our experience
b. Knowledge of time that we gather from books, specially the history books
c. Knowledge of time through unquestionable conventional belief (various kinds of narratives)

These three factors contribute immensely in establishing the knowledge of time for every individual. I have been talking about individuals time and again in the paper as there seems to be a vast difference in the degree and range of knowledge of time amongst individuals. Despite this attested differences in the range and degree of knowledge of time, an amazing commonality is also discovered when we talk about any big/established/famous/memorable event e.g. 1947, 1945 in case of Indian and Korean history respectively, 11th Sept, 2001 in case of American's life and 1914 or 1948 for the world.

If a verb denotes action and action seems to create events on time-scale, an individual must witness a lot of actions and therefore events almost every day. We might have a genuine concern at this point of talk and ask a question as to whether all actions responsible to make events in our life and do we have to store all these events in our mind from the point of view of its referentiality in the time-scale. I would say no in order to answer this question, because the human brain simply cannot bear the load of such a huge task of indexing the information. We must think about the capacity of human mind with regard to the load of memorization that it can bear. If all actions become events and they have to be the part of our knowledge of time, we would need a voluminous brain. So, the human mind has to differentiate and filter out amongst the events that take place every day and those which are far more important and do not happen every day. I would have loved to explain the process through which the brain marks this difference, but this will take us into all together a new field of investigation. I, therefore, do not intend to disgrace the audience and pull them into a whole new area of psycholinguistics where a complicated function of brain, its relationship to the language and other concepts are established. So, let me concentrate on the topic and work out the relationship that exists between time and tense.

There seems to a fundamental advantage to have the knowledge of present-time because this gives us a chance to know other references of time without any

dispute. The knowledge of present time is a self-attested knowledge and it does not require any proof or justification for its existence (in Indian Grammatical Tradition, it is called a 'pramanit-satta'). So, we could draw a time line as follows:

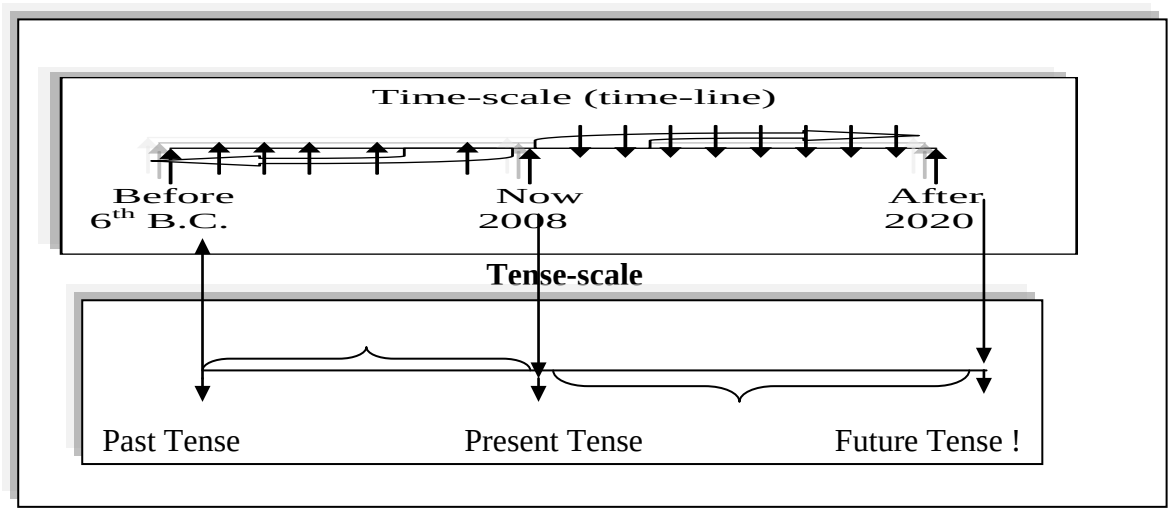


This time scale helps us to establish the knowledge of time of **now**, and since the point '*now*' does not need any proof for its existence, it can function as **a reference point** for the establishment of other points of references of time such as **before** and **after** which can facilitate us to indicate different kinds of actions that we want to incorporate in our communication system.

An example as to how this system of time-scale works, will solve many queries that might be bothering the mind of the audience. Suppose we ask someone, 'how old is s/he?' And when we get to know that s/he is 28 years old, it takes us hardly a second or two to get to know this that s/he is born in 1980. If I ask you a question now as to how do you know that s/he is born in 1980, as s/he did not tell you this. In the answer you would be irritatingly saying, 'Can't you know this, it is so simple. I am sure it is simple but I want you to be conscious as to what did you do to get the year s/he is born. You seem to have subtracted 28 years from NOW, and you got the answer. This is what I meant when I said that the notion of the present time i.e. now does not need any proof and it is a self-proven fact and everyone agrees to this. Not only this, but it also helps us to find out other point of references and thus we know past and future time of references.

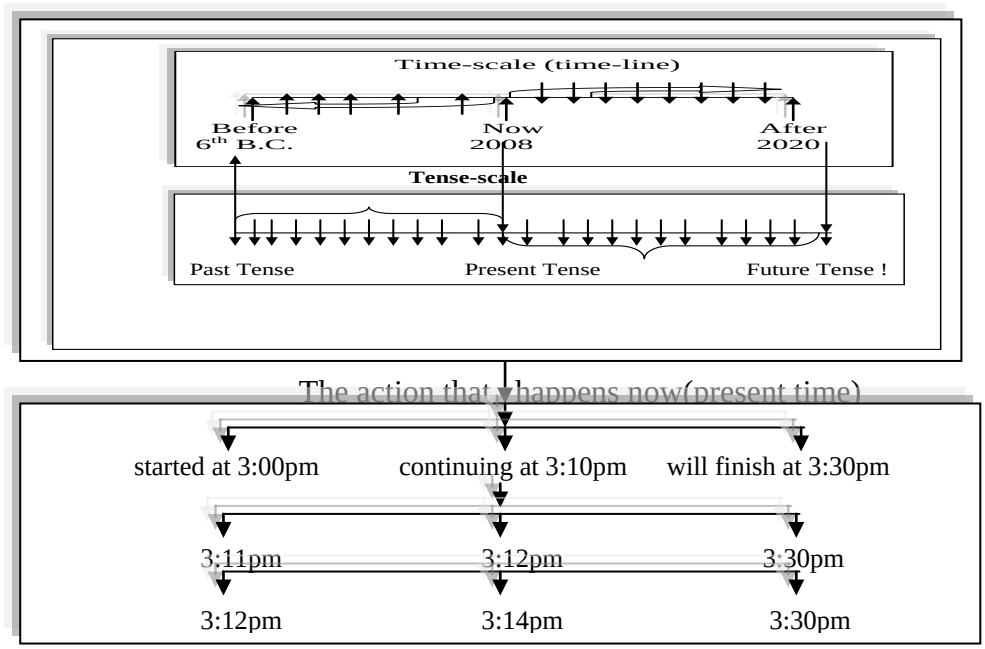
If we turn our attention now to the allocation of actions onto different points of references of time and see how the term 'tense' is conceptualized out of the combination of time and action on the time-scale, we would come to the following diagram:

The formation of tense-scale with the help of that of the time:



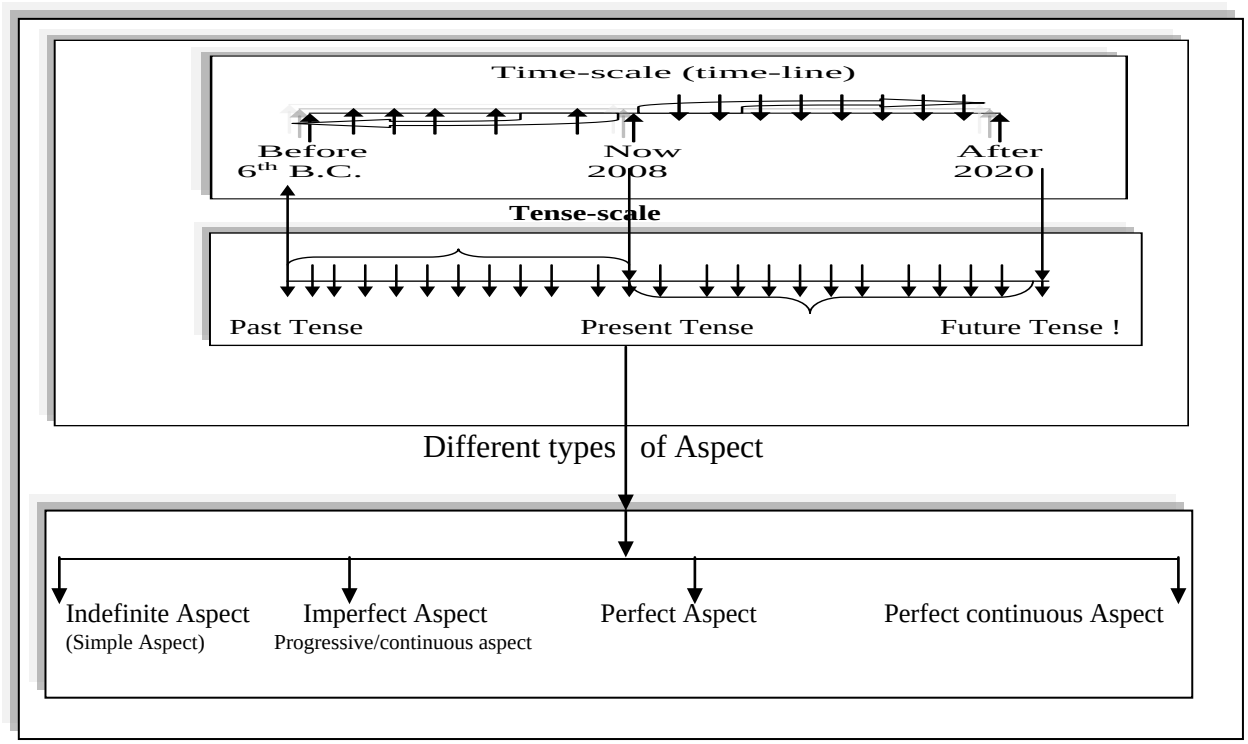
The above diagram explains how we get the notion of 'tense' formed by plotting or allocating the time and action together. In other words, if we know WHEN an action is done and thus the time of the action is known, we can conveniently talk about different tenses with the help of the reference of time. The present tense is about an action that is done at the reference point of time termed as NOW. However, actions that are done during the period of BEFORE at the time reference of ***now*** must fall under the Past tense. Finally, those actions that are planned to be done during the period of AFTER from the time reference of ***now***, is called Future tense.

There are some theoretical as well as pedagogical problems with this Anglo-centric classification of tenses but I will keep those issues aside at the moment as this will divert our attention from what we intended to do here. This classification of tense and its relationship with time seems convincing at the macro level i.e. locating an action in the time-scale and facilitates us to understand as to how the concept of time is imbedded into the explanation of tense as a grammatical category. However, there seems to be problems if we examine the concept of time at a micro level of the distribution of actions denoted by the verb at a given point of time. Look at the following diagram to understand what I intend to say:



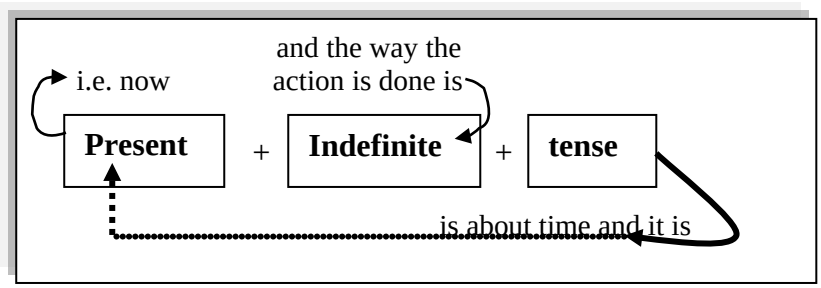
An explanation of the diagram is that we might get trapped into the web of circularity in case we examine the action that happens in present time i.e. at the time of speaking itself. For example, if I talk about my presentation of the day, I would say that I start-**ed** doing my presentation *today* at 3:00pm, I am do-**ing** my presentation *now* at 3:10pm and I **will** finish my presentation at 3:30pm. Now, this seems to be bit problematic, because if you pay attention, you would realize that I have used all three tenses in my statement in talking about the actions that take place in present time. Moreover, by the time I finish saying this, the time 3:11pm is again past and 3:12pm is present and 3:30pm still remains future and so forth.

It is for this reason that the grammarians have introduced a new term called ASPECT. If tense tells us as to when the action is done, aspect tells about how or in what way(s) the action is done. Aspect is of four different types i.e. four different ways of doing an action. Let us see the following diagram to make sense what it means to say that there are four different aspects:



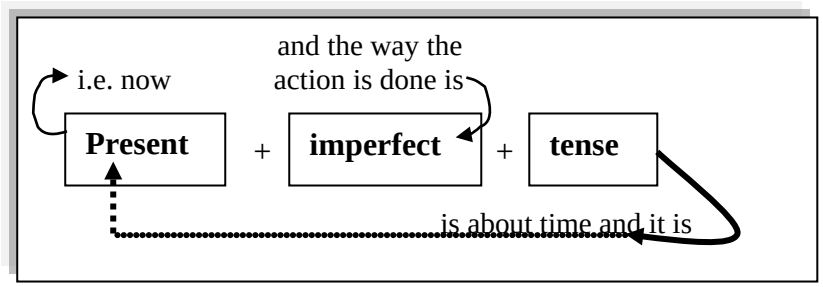
With this diagram of aspect, the notion of time, tense and aspects seems to be completed now. However, it is but necessary to provide with some explanation as to how this whole system works. Let us see the following description for different time, tense and aspects:

1. Present indefinite tense:



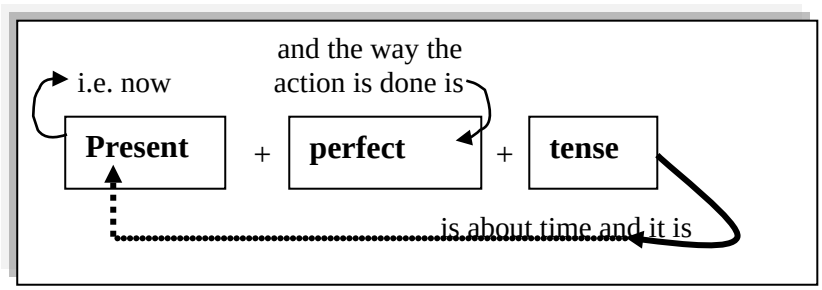
The category tense tells us about a combination of time and action and if we try to find out as to when the action is done or at what time the action is done, the label present in the string tells the computer that it is now i.e. at the time of speaking. The remaining term 'indefinite' in the string will tell the computer that the way in which the action is done is indefinite i.e. there is no definite answer as to whether the action is finished or not. In fact, other than stating that the action takes place in present time of speaking, no other information is supplied with this aspect. Thus, in an example, 'I go to school', there is no information with regard to the verb (action) whether I reached school or not.

2. Present imperfect tense:



The second aspect which is known as ‘present imperfect/continuous/progressive tense’ could be explained in this way. The tense would tell us about the time and we have to find out when the action takes place. The term ‘present’ will tell us about the time i.e. now or at the time of speaking, and the term which remains in the string ‘imperfect’ will tell us the way in which the action takes place. This aspect category is better than the first one, because this informs us better than the first one. The term imperfect/continuous/progressive tells us that the action is not yet perfected and it is in progression at the time of speaking. So, when we say, ‘The boy is writing a letter’, the action described by the verb is witnessed to be in progression.

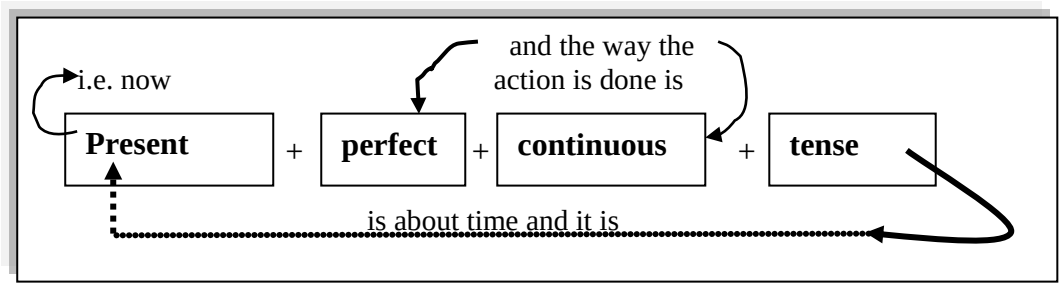
3. Present perfect tense:



This category of tense seems to be the best one in terms of providing the information about the way in which the action is done. Most of the things that we talked about earlier happen in the similar way. The tense would tell us about the time and for time we need to go to the term ‘present’ and this will tell us that it is ‘now’ or at the time of speaking, and the aspect marker ‘perfect’ will describe as to in what way the action is done. There is an important point about perfect aspect that we must explain here. It is not necessary that in ‘present perfect aspect’, the action has to be done always at the time of speaking. The logic is that the effect of the action must prevail in present time or at the time of speaking. Let us explain this with an example. Suppose, a teacher asks his/her students, ‘Have you done your homework?’ The students reply, ‘Yes, we have done our homework’. The issue is that no one did the homework at the time of speaking, but all of them said that they have

done the homework. What does this mean? As I mentioned earlier that it is not necessary that the action has to be done or finished just before the time of speaking with regard to the 'present perfect tense' rather what is necessary is that the effect of the action must prevail in the present time or at the time of speaking. Similarly, if you are offered a lunch and you say, 'I have taken my lunch' does not mean that you are taking lunch or have taken lunch in front of the person who asked the question, but the effect of your taking lunch still prevails in the present time.

4. Present perfect continuous tense:



In this case too, the term *tense* would tell us about the time and time factor would be explained by the word present i.e. now. The way in which is the action is done in this category is called 'perfect continuous'. The 'perfect continuous aspect' is an interesting aspect in English and also in some other languages. It seems that there are some actions such as 'writing a book, teaching a course, and building a house etc.' are some of the actions which would demand not one but a combination of two aspects in order to explain the way the action is performed. The actions mentioned earlier are such that cannot be finished in one given time, neither can it be told in any progressive way. It is therefore, we have a combination of two aspects in English which will explain that the action is started and finished at a given point of time but it would again continue later as well.

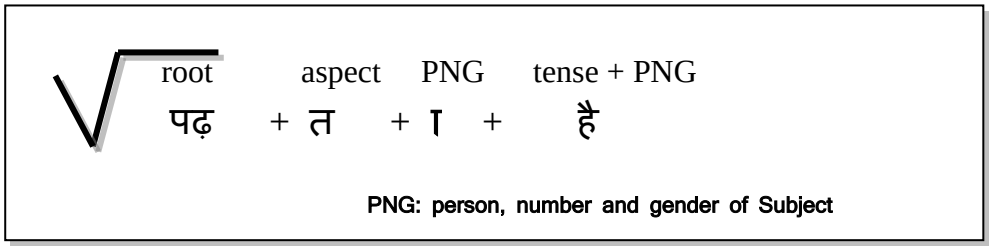
Many languages of the world do not realize or categorize the action in this way and thus would manage with the imperfect/continuous aspect itself. Therefore, instead of four aspects, many languages have only three aspects. We, however, would keep this category as well for the purpose of computational processes in our parser. It is always possible to formulate a sentence for 'perfect continuous tense' in many Indian languages. However, the parser would have the capability to ignore this aspect if in case the language does not mark this aspect morphologically.

It is but obvious that the computer does not have to be strained for these aspects in different tenses. In other words, if we can compute the value of different aspects in one

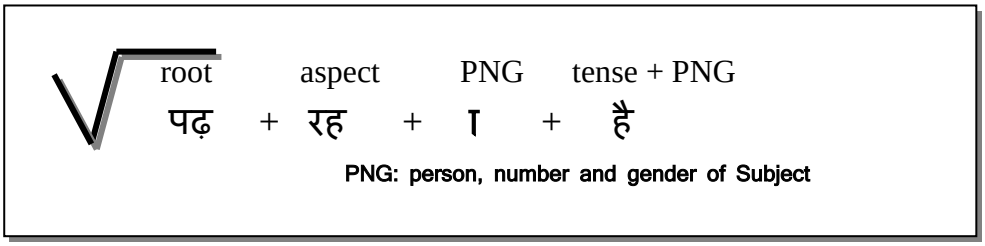
tense, we do not have to keep repeating the process of computation for other tenses because the parser has been given the value of tenses and it automatically recognizes the tense value for past and future. It is for this reason that I do not want to explain these aspects in different tenses.

The computation of Hindi stems for different aspect:

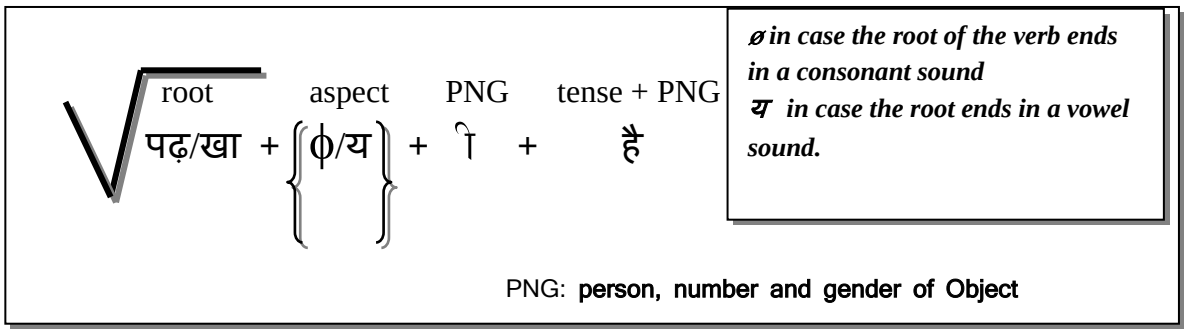
1. Present indefinite tense: लड़का किताब पढ़ता है | (The boy reads the book)



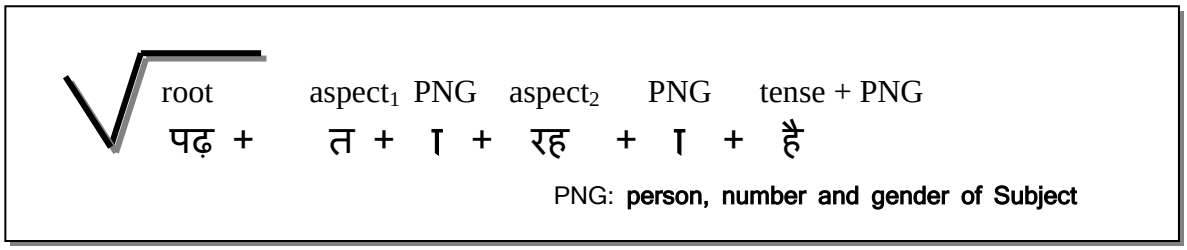
2. Present imperfect tense: लड़का किताब पढ़ रहा है | (The boy is reading the book)



3. Present perfect tense: लड़के ने किताब पढ़ी है or लड़के ने रोटी खायी है
The boy has read the book. or The boy has eaten the bread.



4. Present perfect continuous tense: लड़का सुबह से किताब पढ़ता रहा है |
The boy has been reading the book since morning.



Conclusion:

I would conclude in a line saying that the team who is trying to devise/develop the parser needs your support and best wishes. I believe that the parser has the power to do wonder in terms of the computational value. If we succeed in what we have ventured for, it would prove like a boon for those who just know how to write in English and if they wish to convert their documents in Hindi, they would be able to do it in no time.

Reference:

Bharati, A., V. Chaitanya and R. Sangal (1995) *Natural Language Processing: A Paninian Perspective*, Prentice-Hall of India, New Delhi.

Bharati, A and R. Sangal (1993) Parsing Free Word Order Languages in Paninian Framework, *Proceeding of ACL:93.*

Bloomfield, L. (1949) *Language*, The University of Chicago Press, Chicago.

Carroll, J.M. (2000) *Making Use: Scenario-Based Design of Human-Computer Interactions*, The MIT Press, Massachusetts.

Chafe, W. (1970) *Meaning and the Structure of Language*, The University of Chicago Press, Chicago.

Chomsky, N. (1965) *Aspects of the Theory of syntax*, The MIT Press, Cambridge, Massachusetts.

Das, P.K. (2006) *Grammatical Agreement in Hindi-Urdu and its major varieties*, Lincom-Europa, München.

Grishman, R. (1986) *Computational Linguistics: An Introduction*, Cambridge University Press, Cambridge.

Jesperson, O. (1934) *The Philosophy of Grammar*, George Allen & Unwind Ltd., London.

Kachru, Y. (1966) *An Introduction to Hindi Syntax*, Department of Linguistics, University of Illinois, Urbana, Illinois.

Moravcsik E.A. (1978) 'Agreement' in J.H. Greenberg (ed.) *Universals of Human Language*, Vol.4, Stanford University Press, Stanford.