Information Coding in a language: Some insights from *Pāṇinian* Grammar

Akshar Bharati and Amba Kulkarni

Department of Sanskrit Studies
University of Hyderabad
Hyderabad, India
apksh@uohyd.ernet.in

Abstract. The knowledge of how a language codes information, how much information it codes and where it codes the information is very crucial for a computational linguist working in the area of Natural Language Processing and in particular Machine Translation.

Pāṇini has given utmost importance to the 'information coding' in a language string. This is evident from the use of the same marker *Ṇ* twice making the *pratyāhāra*s apparently ambiguous. *Patañjali* in his commentary points out how one can disambiguate the two *pratyāhāra*s *aṆ* and *iṆ* using the context and finally comments "*vyākhyānato viśeṣa pratipattiḥ na hi sandehāt alakṣaṇam*".

In support of our claim that *Pāṇini* had 'information coding' at the back of his mind while writing the grammar for Sanskrit in the form of *Aṣṭādhyāyī*, we discuss as representatives the 3 *sūtras*: *anabhihite (3.1.1)*, *svatantraḥ kartā (1.4.54)* and *samānakartṛkayoḥ pūrvakāle (3.4.21)*. These 3 sutras precisely point out where the information is coded, how much information is coded and the manner in which the information is coded in Sanskrit.

Finally we seek answers for these three questions in the context of modern English language to arrive at the English grammar from *Pāṇinian* perspective.
Key Words: Information Coding in *Pāṇini*, *pratyāhāra*, *kartā*, agent, implicit coding, Subject, Subject Position, Focus

1 Introduction

Language is a means of communication. A language string encodes the thoughts of a speaker and presents it to the listener. When the listener, on hearing what speaker has uttered, "understands" it, communication is said to take place. What hearer "understands" may not be exactly the same as what the speaker "intended" to say, because language string does not code "completely" what speaker intends to say. It is the shared knowledge about the domain and many other extra linguistic factors which supply the additional information and therefore normally a remarkably good communication takes place.

One may view this whole process of communication as coding and decoding of information using language strings. The difficulty in coding level arises because of the constraints language imposes both at the word level as well as the sentence level. The concepts - independent of language - are mental constructs. They form an open class. Words in any language on the other hand are denumerably finite. This poses a major problem in selection of proper words for communication. Further language has its own grammar rules for combining the words. These grammar rules restrict a speaker to construct his thoughts in a specific manner. Thus a speaker may have to 'fit' his thoughts in a given framework. The lanuguages in general are incommensurate to convey the concepts, much like the rational number system is incommensurate to express an irrational number such as pi. In practice one chooses the best appropriate approximation. Similarly in the world of language communication, one chooses the best possible approximation to express one's thoughts. Moreover, there is always a trade-off between brevity and precision and many a times brevity wins over precision. Thus usually what is coded in a language string is only an approximate representation of the actual thought in the speaker's mind.

Computer being an information processing device, it has been used in the field of Natural Language Processing (NLP) since its invention. There have been efforts to build automatic Machine Translation (MT) systems since 1950s. In the earlier years the difficulties were largely underestimated and the NLP researchers were as enthusiastic as the scientists claiming to build perpetual motion machines in the late 18^{th} century. However the advancements in the computational linguistics(CL), emerging statistical techniques in NLP, advancements in the computer hardware did not wane their enthusiasm and we see thousands of researchers working in the emerging areas of NLP, CL, cognitive science, leading to better theories and better tools for language analysis.

Since languages differ in coding the information, for a person working in the field of NLP, it is appropriate to ask the questions:

1. Where does a language code information? or what means does a language employ to code the information?
2. How much or what kind of relations are encoded in a sentence string? and finally
3. How are the relations coded – explicitly or implicitly?

Such an inquiry will help in knowing the complexity of the task involved and also the upper bounds if any of what can be achieved. It will help in channelising our energies for the achievable tasks rather than diverting our energies in trying to do something unachievable. To be specific, an answer to the first question helps us in designing the parsing strategy. An answer to the second question helps in deciding the level of semantic analysis one can carry out using only the language strings without resorting to the world knowledge. The answer to the final question tells more about the language conventions. Different languages may have different conventions and this may lead to catastrophy in MT if not handled properly.

India has around 2500 years of rich heritage in the linguistic studies. Out of the six *vedāṅga*s (fields of studies necessary to study the vedas) viz. *śikṣā, vyākaraṇa, chanda, nirukta, jyotiṣa* and *kalpa*, the first four are concerned with the language studies. *śikṣā* deals with the pronounciation, *vyākaraṇa* with the grammatical aspects, *chanda* with the prosody and *nirukta* with the etymology. Though all these are important aspects of linguistics, it is the *vyākaraṇa* and the *nirukta* which have major role to play in understanding how a language communicates thoughts from one human being to the other.

Pāṇini consolidated all the earlier grammars for Sanskrit and presented a concise and almost exhaustive descriptive coverage of the then prevalant Sanskrit language. This grammar is in the form of aphorisms – around 4000 divided into 8 chapters of 4 sections each. As Kiparsky(2002) puts it "*Pāṇini*'s grammar is universally admired for its insightful analysis of Sanskrit". Further, though *Pāṇini* wrote the grammar basically for Sanskrit, it provides many ingenious concepts for language analysis, which are 'universal' in nature.

"The goal of Pāṇinian enterprise is to construct a theory of human communication using natural language" (Bharati, 1994). Pāṇinian Grammar(PG), as any other grammar formalism would give, gives a very good theory to identify the relations among words in a sentence. Importance of PG lies in the minute observations of *Pāṇini* regarding the information coding in a language.

In the next section we establish our claim that *Pāṇini* was an information scientist, by citing an example from the *Māhes̀varasūtra*s. The third section discusses three *sūtra*s from *Aṣṭādhyāyī* and show how they answer important questions related to the information coding. In the fourth section, we summarise how these insights have helped us to analyse English grammar from *Pāṇinian* perspective.

2 Repetition of 'Ṇ' in *Māhes̀varasūtra*s

The *Māhes̀varasūtras* form an integral part of the *Aṣṭādhyāyī*. It consists of 14 *sūtra*s. Each *sūtra* has one or more phonological segments terminated by an '*anubandha*' or '*it*'. *Pāṇini* has used around 42 different subsets of phonemes in *Aṣṭādhyāyī*. *Māhes̀varasūtra*s is a linear arrangement of these 42 partially ordered sets(known as *pratyāhāra*s) with markers placed in between indicating different set boundaries(or the end of each *sūtra*). The linear arrangement with markers helps one to obtain the 42 sets by a mechanical procedure thereby facilitating an easy memorisation of these sets. Kiparsky(1994) and Petersen(2004) have given respectively linguistic insight and mathematical proof of the optimality of the *Māhes̀varasūtras* with respect to the placement of the markers as well as the number of markers. Petersen has elegantly shown why the repetition of ‘h’ in the *sūtra*s is necessary and that the choice of ‘h’ is optimal.

Pāṇini has used the same consonant ‘Ṇ’ as an *anubandha* at two different places in the *Māhes̀varasūtra*s. We try to seek a reason behind the use of Ṇ twice. Here are the first 6 *Māhes̀varsūtra*s with repeated ‘Ṇ’.

a i u Ṇ
ṛ ḷ K
e o Ṅ
ai au C
h y v r T
l Ṇ

This makes the *pratyāhāra* ‘aṆ’ and ‘iṆ’ ambiguous since ‘aṆ’ may refer to {a i u} or {a i u ṛ ḷ e o ai au h y v r l}, and the ‘iṆ’ may refer to {i u } or {i u ṛ ḷ e o ai au h y v r l}. *Patañjali* examines all the *sūtra*s that use ‘aṆ’ and ‘iṆ’ and finally concludes that in each of these cases one can resolve the ambiguity. *Bartṛhari*’s commentary on *Mahābhāṣya - dīpikā* is worth mentioning. *Bhartṛhari* observes that[1] the *sāmarthya* (ability to convey a specific meaning), *prasiddhi* (frequency of usage), *liṅga* (indicator) and the *lāghava* (economy) are the deciding factors for resolving the ambiguity arising because of the repetition of ‘Ṇ’.

Aṣṭādhyayī has 5 *sūtra*s which use ‘aṆ’ *pratyāhāra*. They are

ḍhralope pūrvasya dīrghaḥ aṇaḥ (6.3.110)[2]
ke aṇaḥ (aṅgasya hrasvaḥ)[3]) (7.4.13)
aṇaḥ apragṛhyasya anunāsikaḥ (vā) (8.4.56)
uraṇ raparaḥ (1.1.50)
aṇudit savarṇasya ca apratyayaḥ (1.1.68)

In what follows we show how in each of these cases ambiguity can be resolved.

2.1 *Sāmarthya*(ability to convey proper meaning)

The first 3 cases viz.

[1] *ayam ṇakāro dviranubadhyate. atra prakaraṇe ṣatprakārāḥ upasthitāḥ – āsattiḥ vyāptiḥ sāmarthyam prasiddhi liṅga lāghavamiti*

[2] The number 6.3.110 indicates 110th *sūtra* in the 6th *adhyāya*(chapter) and 3rd *pāda*(part).

[3] The words in the brackets are part of *anuvṛtti* (repetition of words from earlier *sūtras*

ḍhralope pūrvasya dīrghaḥ aṇaḥ (6.3.110)
ke aṇaḥ (aṅgasya hrasvaḥ) (7.4.13)
and
aṇaḥ apragṛhyasya anunāsikaḥ (vā) (8.4.56)

contain the words '*hrasvaḥ*', '*dīrghaḥ*' and '*pragṛhya*'. '*hrasva*' and '*dīrgha*' are the properties of vowels and only a vowel can get the technical name (*sañjā*) - *pragṛhya.* Or, in other words, there will never be cases where '*hrasva*', '*dīrgha*' and '*pragṛhya*' will qualify any of the phonemes from {h y v r l}. Therefore *Patañjali* argues that if in these three *sūtra*s 'Ṇ' refers to the 2nd Ṇ in the *sūtra*s, since the rules are not applicable in cases of {h y v r l}, it would have been sufficient to use the *pratyāhāra* 'aC'(set of vowels) (which is already in use and hence does not lead to a new *pratyāhāra* also), thereby resorting to economy (*lāghava*). In fact further from the point of view of economy, he argues that, even 'aC' need not be mentioned, being the default case, leading to further *lāghava* at *sūtra* level. However the fact that *Pāṇini* has mentioned 'aṆ', he meant 'aṆ' referring to the smaller set {a i u} and not the bigger one. Thus it is the words - '*hrasva*', '*dīrgha*' and '*pragṛhya*' in the context which facilitate the word 'aṆ' to convey one meaning over the other. *Bhartṛhari* terms this as '*sāmarthya*' - an ability of a particular meaning to express itself (in a particular context).

2.2 *Prasiddhi*(frequency of usage)

In the next *sūtra* '*uraṇ raparaḥ*' (1.1.50), the possibility of 2nd 'Ṇ' is ruled out on the basis of unavailibility of any example which involved the bigger set {a i u ṛ ḷ e o ai au h y v r l}. *Patañjali* discusses two examples in his commentary and he points out that either the effect of the rule is further nullified by other *sūtra* or application of this *sūtra* leads to redundancy in some other *sūtra* which is undesirable and hence concludes that if at all *Pāṇini* meant 2nd Ṇ, he could have used a smaller *pratyāhāra* 'aC'. Since *Pāṇini* used 'aṆ' and there is no evidence otherwise, *Patañjali* concludes that 'Ṇ' in this *sūtra* is the 1st one (because in all the previous *sūtra*s involving 'aṆ', it is the 1st 'aṆ' which is being used) and not the 2nd one. According to *Bhartṛhari*, it is the *prasiddhi* (frequency of usage) which is the deciding factor in this *sūtra.*

2.3 *Liṅga*(marker)

The 5th *sūtra* that uses 'aṆ' is

aṇudit savarṇasya ca apratyayaḥ (1.1.68)

From this *sūtra* alone it is not obvious which 'aṆ' is meant. There is another *sūtra* '*u ṛt (7.4.7)*' which says 'ṛ' becomes 'ṛt'. The 't' in 'ṛt' makes 'ṛ' *tapara* which means the 'ṛ' represents itself and not its *savarṇa*[4]. If the 'ṇ' in the *pratyāhāra*

[4] refer the *taparastatkālasya* – 1.1.69

'*aṇ*' were the first 'ṇ', it would not have been necessary to mark '*ṛ*' as *ṛt.* The very presence of the *sūtra* 7.4.7 therefore indicates that 'ṛ' is member of the 'aṆ' in 1.1.68 and hence the 'Ṇ' in 1.1.68 is the 2$^{\text{nd}}$ 'Ṇ'.

2.4 *Lāghava*(economy)

Finally in case of 'iṆ', it is observed that if *Pāṇini* wanted to mention the 1$^{\text{st}}$ 'Ṇ', only two phonemes 'i' and 'u' being involved, he used '*yvoḥ*' instead of '*iṆaḥ*'. In fact '*yvoḥ = y v o ḥ*' involves 3.5 (=0.5 + 0.5 + 2 + 0.5) *mātrā*s (time measure of utterence of a phoneme) whereas '*iṇaḥ = i ṇ a ḥ*' involves 3 (= 1 + 0.5 + 1 + 0.5) *mātrās.* Thus in spite of 'non-economy' (*gaurava*) of 0.5 *mātrā, Pāṇini* prefers '*yvoḥ*' over '*iṇaḥ*', naturally for the purpose of '*lāghava*'(economy) in other cases.

2.5 Why repetition?

Patañjali at the end of the discussion on this topic in *Mahābhāṣya* raises a valid question - was there a dearth of consonants that *Pāṇini* used the same phoneme twice? In response he warns

> '*vyākhyānataḥ viśeṣa pratipattiḥ na hi sandehāt alakṣaṇam*'

(if one can not resolve the ambiguities, one should not jump to the conclusion that the *sūtra*s are defective.)

Had *Pāṇini* used some other consonant as an *anubandha*, he would have lost the opportunity to train the students in paying attention to the different means of information coding in a sentence.

Should we then not conclude that *Pāṇini* was aware of the ambiguities a natural language has and wanted to train the students of *vyākaraṇa* to pay attention to different sources of information available for disambiguation? And that he uses the very first opportunity to train the students - right from the *Māheṣvarasūtra*s with which the study of *Aṣṭādhyāyī* commences?

Though a substantial part of *Aṣṭādhyāyī* deals with the rules related to morphology, an important section of it deals with concepts important from the language analysis point of view. Two of the important sections are those related to *kāraka* and *samāsa.* It is the *kāraka* - *vibhakti* mapping which provides the bridge between the semantics and syntax. In this section, we show with examples, the importance *Pāṇini* has given to the information coding in a language string.

3 Dynamics of Information coding in Sanskrit

Before we proceed further a brief introduction of *Pāṇinian* Grammar is in order. The underlying axioms of the *Pāṇinian* Grammar are:

1. Each word consists of two parts: a root(stem) and a (primary) suffix. A root can be a lexical item available in the *dhātupāṭha*, a nominal stem, a derived nominal or a derived verbal stem. The (primary) suffixes are of two types: nominal(*sup*) and verbal(*tiṅ*). In addition there are derivational suffixes which produce new stems. There are 4 major ways of deriving new stems from the nominal or the verbal stems. They are:
   (a) adding nonfinite verbal suffix (*kṛt*),
   (b) adding a *taddhita* suffix to derive a new nominal base,
   (c) deriving *samāsa*s (compound nouns) (there are 6 ways of doing this) and
   (d) adding special verbal/nominal suffixes (*sanādi*) to derive new nominal/verbal bases.
   Meaning in each of these 4 cases is compositional (but for some exceptional cases of compounds).

2. The primary suffixes mark relations between words. These are of two types: *kāraka* relations and *kāraketara* (other than *kāraka*) relations. *Samānādhikaraṇa, tādarthya* (purpose), *hetu* (cause), *sambandha-ṣaṣṭhī* are some of the *kāraketara* relation. A relation between a noun and a verb is expressed in terms of *kāraka*s. A *samānādhikaraṇa* relation marks a relation between an adjective and a noun or a noun with its apposition (Joshi, 1998; vol VII). The concept of *kāraka*[5] plays the central role in *Pāṇinian* grammar. A verb denotes an action. Various participants involved in this action have different roles – named as '*kāraka*' roles. There are six '*kāraka*' roles: *kartā, karma, karaṇa, sampradāna, apādāna* and *adhikaraṇa.* (One thematic role may map to different *kāraka* roles and one *kāraka* role may stand for different thematic roles in different contexts.)

3. The action denoted by a verb typically stands for a complex activity which may further be split into subactivities. For example, the activity corresponding to the action 'opening of a lock' consists of the following subactions(Bharati, 1994):
   (a) a person inserting a key into the lock,
   (b) the pressing of levers and moving them by a key,and
   (c) moving of the latch and opening of the lock.
   The mapping of the semantic roles to the *kāraka* roles depends on how the speaker views the activity (*vaktṛ vivakṣā*). Speaker may choose to focus on the subactivity of the key, in which case the key will be a *kartā* rather than a *karaṇa.*

[5] *kāraka* and '*kartā*','*karma*', etc. are the technical terms which *Pāṇini* uses. These technical terms being self-explanatory, he does not define them formally but he indicates their intended meanings whenever necessary.

```
---  semantic level (what the speaker has in mind)
 |
 |
 |  ---  Thematic roles
 |
---  karaka level
 |
 |
---  vibhakti level
 |
 |
---  surface level (uttered sentence)
```

Fig. 1. Levels in the Paninian model

4. Pāṇinian grammar visualises the whole process of sentence generation (see fig. 1) as a 3 step (or a 4 level) process(Bharati 1994). The first step involves mapping the concepts into words and assigning proper *kāraka* roles to different nouns representing various concepts. Choice of voice along with a verb decides the case markers different nouns will take based on their *kāraka* roles. To a large extent this is a mechanical step and needs higher level semantics very rarely. In the last step the morphophonemic rules then take over and a sentence is generated.

   These relations, "serve as intermediaries between grammatical expressions and their semantics"(Cardona,1978) providing a bridge between the surface form and its meaning.

   When we look at the way *Pāṇini* has described the Sanskrit language, it is very clear that he paid utmost attention to the information coding in a language string. In support of our claim we produce 3 evidences from *Pāṇini*'s *Aṣṭādhyāyī* where *Pāṇini* makes subtle observations about the information coding in a sentence.

3.1 'Where' is the information coded

Look at the sentences:

San: *rāmaḥ grāmam gacchati.*
Eng gloss: Rama{nom} village{acc} go{active_voice,pr_tense,3_person,sg}
and

San: *rāmeṇa grāmaḥ gamyate.*
Eng gloss: Rama{instr} village{nom} go{passive_voice,pr_tense,3_person,sg}

In Sanskrit, it is the suffixes which code the information about relations between words. Hence, from the above sentences, one would say, in case of active voice, the nominative case of a noun marks it as a *kartā* and the accusative case of a noun marks it as a *karma*. In case of passive voice, *kartā* gets an instrument case and *karma*(in case of transitive verbs) gets a nominative case. We further note that the *kartā*(*karma*) and the verbal suffix agree in number and person in active(passive) voice. Moreover Sanskrit also allows pro-drop. '*gacchāmi* (go,pr_tense,first_per,sg)' is a perfect and complete sentence. How to account for such pro-drop cases? So we require separate rules to account for the agreement and the pro-drop. *Pāṇini* neither derives passives from actives nor does he have special treatment for pro-drop.

Pāṇini handled all these cases in a very compact and elegant way. He says

1. *laḥ karmaṇi ca bhāve ca akarmakebhyāḥ (kartari) 3.4.69*
2. *anabhihite 3.1.1*
3. *kartṛkaraṇayoḥ tṛtīyā 2.3.18*
4. *karmaṇī dvitīyā 2.3.2*
5. *prātipadikārthaliṅgaparimāṇavacanamātre prathamā 2.3.46*

Let us try to understand why he framed the rules this way?

Let 'W_1 W_2 V' be a sentence, where W_1 and W_2 are the nouns and V is a verb. Further let

$W_1 = R_1 + S_1$,
$W_2 = R_2 + S_2$ and
$V = R_3 + VS$,
where R_1, R_2 and R_3 are the roots and S_1, S_2 and VS are the suffixes.

We know that R_1, R_2 and R_3 relate to the real world, expressing the concepts. It is S_1, S_2 and VS which relate the three words W_1, W_2 and V with each other. Following the assumption that the participants in the action denoted by the verb are related to the verb, is it the suffix S_1 (or S_2) which relates W_1 (or W_2) with the V, or is it S_1(or S_2) and VS together mark the relation, or is it VS alone which marks the relation? In case of S_i denoting non-nominative cases it is very clear that it marks the relation between W_i and V. But in case of a nominative case, there is a problem. If we assume that in case of a nominative case also, S_i marks the relation between W_i and V, then how can one account for the sentences, where there is no S_i as in pro-drop cases such as '*gacchāmi*'? Naturally, one has to resort to 'VS' as a relation marker. But any relation should have two relata. In this example, there is only one word. Then how does one account for the missing relatum? Pro-drop in Sanskrit is possible only in first person and second person. Since the first person and second person pronouns are unique, even if they are not mentioned, by default they are understood. There is no loss of information. However, one can not drop the relatum in case of third person in Sanskrit. Thus we see that, it is 'VS' which marks the relation between W_i and V, in case of nominative case. This is what has been observed by *Pāṇini* when he states

laḥ karmaṇi ca bhāve ca akarmakebhyāḥ (kartari) 3.4.69

It is the *lakāra* (tense-aspect-modality marker) which expresses the *kartā*, *karma* or *bhāva*(action).

This is a very subtle observation which leads to an obvious question: what does then the nominative case signify? *Pāṇini* says

prātipadikārthaliṅgaparimāṇavacanamātre prathamā 2.3.46

The nominative case just indicates the gender, number etc. and not any *kāraka* relation. Having handled the nominatve case, now *Pāṇini* describes further the *kāraka* - *vibhakti* mappings. The section on mapping the *kāraka* relations into *vibhakti*s starts with the *sūtra* '*anabhihite*' (that which has not been expressed).[6] In case the relation has not been expressed by any of the means, then the rules from 2.3.2 to 2.3.73 come into effect and the unexpressed *kāraka* relations are expressed through the *vibhakti*s.

As a consequence of this generalisation, as Kiparsky observes(Kiparsky, 2002), 'Actives, passives, sentences and nominals are alternative realisations of the same underlying relational structure!'.

What we learn from the way *Pāṇini* framed the rules is: to look for 'where the information is coded'. The very fact that language allows pro-drop triggers that it is the verbal suffix which codes the *kāraka* relation and not the nominative case. For someone who is interested in 'understanding' a text in another language, or 'processing' it for some NLP applications it is crucial to know 'where' exactly the language codes the information.

3.2 How much information is coded

In the previous section we saw that the *vibhakti*s (case markers) are determined by the *kāraka* role a noun phrase has with respect to the verb.

$$vibhakti = \mathrm{f}(kāraka,prayoga)$$

[6] *Kātyāyana* in his *vārtika* on this *sūtra* states that there are 4 ways by which the *kāraka* relations can be expressed – by means of *tiṅ* suffix, *kṛt* suffix, *taddhita* suffix (derivational suffix deriving a noun) and *samāsa* (compound).

vibhkati(case marker) and the *prayoga*(voice) are the surface level realities. *kāraka*s are the basic syntactico-semantic categories.
In active voice(*kartari prayoga*), since the verbal suffix expresses the *kartā, kartā* has a nominal case. Therefore in the sentences

1. *rāmaḥ kuñcikayā tālam udghāṭayati.*
2. *kuñcikā tālam udghāṭayati.*
3. *tālaḥ udghāṭyate.*

which are in active voice, *rāmaḥ, kuñcikā* and *tālaḥ* being in nominative case are the *kartā*s. Semantically however, *rāma* is the agent, *kuñcikā* is the instrument and *tālaḥ* is the goal. It is obvious that by calling all these three *kartā*s, the actual semantic roles are not captured and one needs one more mapping from these *kāraka* roles to the thematic roles to arrive at the semantics. Natural question is why *Pāṇini* did not go for the semantic analysis? And why did he choose the *kāraka* level analysis? *Pāṇini* has not written a single comment on the purpose of *aṣṭādhyāyī*, the approach he followed, etc. *Pāṇini* observes

svatantraḥ kartā (1.4.54).

Patañjali in his *Mahābhāṣya* has elaborated on it. An activity involves more than one participants. The underlying verb expresses the complex activity which consists of subactivities of each of the participants involved. For example, in case of opening of lock, three subactivities are very clearly involved(Bharati,1994), viz.

1. the insertion of a key by an agent,
2. pressing of the levers of the lock by an instrument (key), and
3. moving of the latch and opening of the lock.

Though in practice, to a large extent all the three subactivities starting from 1 through 3 together means 'opening of the lock', sometimes the subactivities 2 and 3 together are also referred to as 'opening of a lock' and the activity 3 alone is also referred to as 'opening of the lock'. Different languages may have different lexical items expressing these subactivities. But in case the lexical items are the same, it is ambiguous. When we say *rāma*, *kuñcikā* and *tālaḥ* are the *kartā* of opening of a lock, *rāma* is the *kartā* of the activities 1 through 3, *kuñcikā* that of 2 through 3 and *tālaḥ* that of 3.

Patañjali interprets '*svatantraḥ kartā*' as: in the absence of participants capable of performing subactvities a_j ($j < k$), the participant performing the subactivity a_k will be the *kartā*. (As *Patañjali* puts it, in the absence of a king, the senior most minister will have the powers of king.)

Thus in the absence of an agent(*rāma*), by promoting an instrument (*kuñcikā*) to *kartā*, *Pāṇini* draws our attention to the fact that language does not code information completely. Information related to the semantic encoding is not coded in a language string. To arrive at the conclusion that *kuñcikā* is an instrument and *tālaḥ* is a goal, one has to appeal to the world knowledge. The greatness of *Pāṇini* lies in "identifying exactly how much information is coded and then giving it a semantic interpretation" (*sūtras* 1.4.23 - 1.4.55). This level of semantics is the one which is achievable / reachable through the grammar rules and the language string alone. This puts an upper bound for the analysis, making it very clear what is guaranteed and what is not. We can extract only that which is available in the language string 'without any requirement of additional knowledge'. To give an analogy, you can not use low quality energy to do the high quality work.

3.3 How (manner) is the information coded?

The *sup* and *tiṅ* suffixes assign *kāraka* roles to the nouns. The principles governing the relations between these suffixes with the *kāraka* roles are as under (Kiparsky, 2002).

1. Every *kāraka* must be expressed by a morphological element.
2. No *kāraka* can be expressed by more than one morphological element.
3. Every morphological element must express something.

We have seen earlier that every suffix (*sup* or *tiṅ*), except the nominal case can express only one *kāraka*. Similarly, every *kāraka* can be expressed through one and only one suffix.

Now consider a sentence

San: *rāmaḥ dugdham pītvā śālām gacchati.*
Eng gloss: Rama{nom} milk{acc} after_drink{gerund} school{acc} go{pr,active,3p,sg}

In this sentence, there are two verbs viz. *gam* and *pī*. Both of them have a mandatory expectancy of two *kāraka*s viz. *kartā* and *karma*. Further the relation between the subordinate verb and the main verb should also be marked. Thus there are 5 relations which need to be marked. In the above sentence, there are only 5 words and one of them is in nominative case. Hence only 4 relations can be expressed through the suffixes. Relations that are expressed by the suffixes are shown in figure 2.
The *kartā* of the verb *pī* is not marked explicitly. A native speaker, however, does not have any problem in answering the question 'who drank the milk?'. This indicates that it is the 'Language Convention' that tells: in case of 'ktvā' suffix[7] the *kartā* of the subordinate verb is the same as that of the main verb. *Pāṇini* has postulated this in terms of a *sūtra*

'samānakartṛkayoḥ pūrvakāle' (3.4.21)

[7] which indicates that the action corresponding to the verb with 'ktvā' suffix takes place before the action indicated by the main verb

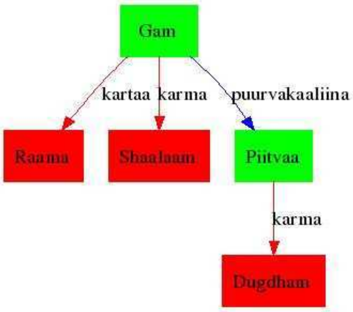

Fig. 2. modifier-modified relations

It is the language convention which gives a license to not to code the information explicitly. The implicit coding of the information may need extra processing for making such a knowledge explicit. It then becomes crucial for MT developers to know what is coded explicitly and what is coded implicitly. If the two languages have different language conventions, one needs to make implicit information explicit in other language. This may lead to unacceptable constructions, or even to a catastrophe, if not handled properly.

Consider
San:*vanāt grāmam adya upetya odanam āśvapatyena apāci.* (Kiparsky, 2002)
Eng gloss: forest{abl} village{acc} today after_reaching rice{nom} Asvapati{inst} cook{passive, past, 3pr,sg}

In Sanskrit, following the *sūtra*, '*samānakartṛkayoḥ pūrvakāle*', it is clear that it is *āśvapati* who returned and it is he who cooked. But such constructions are not allowed in English. English needs passive absolutive, if the finite verb is in passive. Sanskrit uses same '*ktvā*' in both the active as well as passive form of the finite verb. Hence in MT they pose a problem, as they may lead to unacceptable / ungrammatical constructions.

What we learn from this exercise is, *Pāṇini* gave utmost importance to the information coding in a language. And hence we claim that any grammar which is developed with the three questions in mind: **where**, **how much** and **how** is the information coded, would be truly in *Pāṇinian* spirit.

4 English Grammar: A *Pāṇinian* perspective

Here we illustrate with an example of English Grammar, how the above three questions help us in developing *Pāṇinian* Grammar for English.

4.1 Where does English code the information and how much does it code

English codes the crucial information about the noun-verb relations in pre-verbal and post-verbal positions. However we also come across examples where there is a deviation from the normal SVO order as in
Fruits, I like.
or sentences where subject is not in the subject position as in
Never was the sea so calm!
Then the question is if the information about noun-verb relations is coded in position, 'where' exactly are the relations coded and 'how much' of the information is coded. It has been argued(Bharati, 2005) that the missing accusative marker in modern English has been compensated by the Subject Position and further the fact that there is no morpheme expressing the yes-no question marker, this information is coded in the subject-auxiliary inversion. Both these constraints force the subject position to be non-empty.

4.2 Subject in English is *abhihita*

Subject is reported to have the following properties in the English Grammar literature.

1. It occupies the preverbal position.
   example: *Ram* is going to school.
2. It agrees with the finite verb.
   example: *She* go*es* to the market.
3. It is in nominative case.

From these properties one may infer that the subject is '*abhihita*' and it typically occupies the pre-verbal position[8] - called the 'Subject Position'. The *abhihita* is *kartā* (or *karma*) if the verb is in active(passive) voice, as is clear from the following examples:

1. His face dripped with sweat.
2. The wall crawled with roaches.
3. The church echoed with the voices of the choristers.
4. My guitar broke a string.
5. My car burst a tyre.
6. The fifth day saw our departure.
7. The hall has witnessed many historic events.
8. Ravana was killed by Ram.

[8] It is the position immediately to the left of the first auxiliary verb(*avyavahita pūrva*) or to the left of the finite verb in case of absence of auxiliary verbs.

4.3 Subject(*abhihita*) need not be in the Subject Position

Dummy it: The constraint that subject position can not be empty further forces one to use ‘it’ as a filler as in the following cases

It is raining.
It seems John has left the office.

Subject Raising: Since languages prefer brevity, there is a tendency to eliminate the filler element(since it does not carry any meaning) wherever possible. This leads to subject raising phenomenon in English. For example ‘John’ in the sentence

It seems John has left the office.
‘raises’ to the subject position of ‘seem’ replacing the ‘it’ leading to
John seems to have left the office.

Focus: The subject position being the sentence initial position, also serves the purpose of focussing. Hence in order to focus the manner or to express the factuality or happening or existence, the subject moves past the verb and the subject position is occupied either by ‘here / there’ or the manner adverbs as in

Never was the sea so calm!
There entered the hall the charming prince!

***Samānādhikaraṇa* and Object raising:** Sanskrit allows verbless sentences such as ‘*aśvaḥ* (horse) *śvetaḥ*(white)’. But English mandatorily requires a verb. Thus in case of adjectives having a *samānādhikaraṇa* relation with a noun or a verbal expression denoting an activity, English mandatorily requires a ‘be’ verb, as in

She is beautiful.
She is a teacher.
Running is good for health.

When such an activity corresponds to a transitive verb and the ‘subject position’ becomes ‘heavy’, there is a tendency to shift the subject to post-verbal position as in

It is possible that the earth is flat.[9]

[9] compare with ‘That the earth is flat is possible’.

It is tough to believe that University would fire John.[10]

Here again, since 'it' is just a filler, there is a tendency to move 'John' to the subject position leading to

John is tough to believe University would fire.

Appendix - I contains the rules emerging out of this discussion related to positioning of *abhihita* in English expressed in the form of an algorithm. Strictly speaking no special efforts have been taken to write them in the *sūtra*[11] style

4.4 Language convention in English:

The language conventions specific to a language may be discovered by asking a question 'Is the information coded explicitly or implicitly?'. The implicit coding of the information is due to the 'language convention'. Consider the sentence

Mohan dropped the melon and burst.

Though this sentence is not meaningful, still, if a native speaker is forced to answer the question 'who burst?', the only answer one gets is 'Mohan burst'. However 'Mohan' is not in the subject position of 'burst'. Or in other words, the relation of 'Mohan' with 'burst' is not explicitly coded in the sentence. It is the language convention that allows a native speaker to infer this meaning.

We have concentrated only on one phenomenon of English viz. the 'Subject Position', and the information coded in it. This is just a glimpse of how one can approach another language from 'information coding' perspective to discover the grammar. Contrastive study of two languages following this approach also leads to the discovery of parameters in which two languages differ(Bharati, 2005).

5 Conclusion

With the emergence of Linguistics, linguists started recognising the importance of *Pāṇini*'s grammar. And now with the advent of computer technology, computer scientists have started recognising *Pāṇini* as an information scientist.

In this paper we tried

1. to justify the claim of the computer scientists (Huet,2007): '*Pāṇini* is a father of informatics' and

[10] compare with 'That University would fire John is tough to believe.'

[11] *alpākṣaram asandigdham sāravat viśvatomukham*
astobham anavadyam ca sūtram sūtravido viduḥ

2. to show how the information theory related questions help one to write grammar for any language in true *Pāṇinian* spirit.

Acknowledgement The author acknowledges the fruitful discussions she had with Vineet Chaitanya and K. V. Ramkrishnamacharyulu. The author also thanks V Sheeba and Devanand Shukla for helping her in translating the rules in Appendix-I to Sanskrit.

6 Appendix-I

This is a 'patch' to *Pāṇini*'s *aṣṭādhyāyī* to account for 'Subject' and 'Subject position' in English language

1. *āṅgle*
2. *kriyāpada-avyavahita-pūrvam* Subject *sthānam*
3. *asarvanāmasu dvitīyā-adarṣanatvāt*
4. *abhihitam* Subject *sthānakam*
5. *kriyāpade-uddeśye* (Subject *sthāne*) Here *vā* There *iti*
6. *kriyāviśeṣaṇe (uddeśye) kriyāviśeṣaṇam* (Subject *sthāne*)
7. *antarbhūta kartṛtve* it (Subject *sthāne*)
8. Seem *kriyāyām* (Subject *sthāne*) *vā vākyakarma-abhihitam*
9. *vākya-karma-kriyāyām* to infinitive *ādeśaśca*
10. *icchārthakeṣu asamānakartṛkatvam ca*
11. (*icchārthaka-asamānakartṛkeṣu) gauṇa-kriyā-* Subject *dvitīyāyām*

The above patch has been translated into English below.

1. In modern English,
2. pre verbal position is called *'Subject Position'*.
3. In view of absence of 'accusative marker',
4. *abhihita* occupies the 'Subject Position'. (Hence the *abhihita* is also called a 'Subject'.)
5. If the verb is to be focussed, then the Subject Position is occupied by 'there' or 'here'.
   example: Here comes the bus!
   example: There entered the hall a charming prince!.
6. In case the manner of the activity is to be focussed, then the subject position is occupied by the adverb expressing the manner.
   example: Never was the sea so calm.
   example: Uneasy lies the head which wears a crown.
7. If the verb has implicit *kartā*, then the Subject Position is occupied by 'it'.
   example: It rains.
8. If the main verb is 'seem', then optionally the Subject of the 'vākya karma' occupies the Subject Position of the main verb

9. In such cases the verb of the vākya karma assumes 'to infinitive' form.
   example: John seems to have left.
10. In case of 'icchārthaka dhātu', Subject sharing is optional.
    example: I want to go.
    example: I want him to go.
11. (In case of iccārthaka dhātu), if the Subject of the main verb differs from that of the secondary verb, then the Subject of the secondary verb takes an accusative marker.

References

1. Abhyankar and Limaye *Mahābhāṣya - Dīpikā of Bartṛhari* Bhandarkar Oriental Research Institute, 1967
2. Bharati, Akshar, Amba P Kulkarni, *English from Hindi viewpoint: A Paninian perspective,* Platinum Jubilee conference of LSI at HCU, Hyderabad, Dec 6-8, 2005
3. Bharati, Akshar and Rajeev Sangal, A Karaka Based Approach to Parsing of Indian Languages, In *COLING90: Proc. of Int. Conf. on Computational Linguistics (Vol. 3), Helsinki*, Association for Computational Liguistics, NY, August 1990, pp. 25-29.
4. Bharati, Akshar, Rajeev Sangal, and Vineet Chaitanya, Natural Language Processing, Complexity Theory and Logic, In *Foundations of Software Technology and Theoretical Computer Science 10, Lecture Notes in Computer Science 472*, Springer Verlag Berlin, 1990a, pp.410-420.
5. Bharati Akshar, Vineet Chaitanya, Rajeev Sangal, *NLP A Paninian Perspective,* Prentice Hall of India, Delhi,1994
6. Cardona, George, *Panini: A Survey of Research*, Mouton, Hague-Paris, 1978.
7. Cardona, George, *Panini: His Work and Its Tradition, (Vol. 1: Background and Introduction)*, Motilal Banarsidas, Delhi, 1988.
8. Huet Gèrard, Keynote address at the First International Sanskrit Computational Linguistics Symposium, Paris, 2007.
9. Jigyasu, Brahmadatt. *Ashtadhyayi (Bhashya) Prathamavrtti, three volumes*, Ramlal Kapoor Trust Bahalgadh, (Sonepat, Haryana, India), 1979. (In Hindi)
10. Joshi, S.D. (editor). *Patanjali's Vyakarana Mahabhashya*, (several volumes), Univ. of Poona, Pune, 1968.
11. Joshi, S.D. and Roodebergen J.A.F. *The Aṣṭādhyāyīof Pāṇini* (several volumes), Sahitya Akademi, Delhi, 1998
12. Kiparsky, Paul. *On the Architecture of Panini's Grammar*, CIEFL, Hyderabad, Jan 2002
13. Radford, Andrew, *Syntax: A minimalist introduction*, Cambridge University Press, 1997