Analyzing English Phrases from Paninian. inian Perspective

by

Akshar Bharati, Sukhada, Dipti Misra Sharma

in

17th International Conference on Intelligent Text Processing and Computational Linguistics

Konya, Turkey

Report No: IIIT/TR/2016/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2016

Analyzing English Phrases from Pāṇinian Perspective

Akshar Bharati, Sukhada, and Dipti M Sharma

Language Technologies Research Centre,
IIIT Hyderabad, Telangana 500032, India
sukhada@research.iiit.ac.in
dipti@iiit.ac.in

Abstract. This paper explores Pāṇinian Grammar (PG) as an information processing device in terms of 'how', 'how much' and 'where' languages encode information. PG is based on a morphologically rich language, Sanskrit. We apply PG on English and see how the Pāṇinian perspective would deal with it from the information theoretical point of view and its effectiveness in machine translation.

We analyze English phrases defining *sup* (nominal inflections) and *tiṅ* (finite verb inflections) and compare them with the notion of *pada* (an inflected word form) and *samasta-pada* (compound) in Sanskrit.

Sanskrit encodes relations between nouns and adjectives and nouns in apposition through agreement between gender, number and case markers, whereas English encodes them through positions. As a result, constituents are formed. It appears that an English phrase contains more than one *pada*, hence, cannot be similar to a *pada*. However, we show the linguistic similarities between a *pada*, *samasta-pada* and 'phrase'.

1 Introduction

Languages encode linguistic information in terms of explicit markers or positions of the words. This inspires us to analyse how a source language syntax encodes linguistic information, so that it can be transfered to target language. Pāṇinian Grammar (PG) gives insights to explore 'how', 'how much' and 'where' languages encode linguistic information [6]. In this paper, we use concepts from PG and apply them to English and show how it can help in machine translation (MT).

PG analyses a word as a combination of 'root' (*prakṛti*) and 'suffix' (*pratyaya*) [8]. Pāṇini uses the term *pada* for the words that are ready to participate in a sentence. A *pada* contains explicit information about a word's semantic relation with other words in a sentence.

The word derivation process in PG takes a nominal stem and/or a verbal stem as the basic input and terminates the process with the derivation of *pada* [22]. Since a *pada* is formed with respect to an actual sentence structure, it is called the highest derivative and is a syntactic unit rather than a morphological unit in the Pāṇinian system [22].

Sanskrit uses two different terms, *śabda* and *pada*, both of which are roughly translated as 'word' in English. In Sanskrit, *śabda* is used for linguistic expressions ranging from an individual speech sound to an utterance [23], whereas, a *pada* is a primary syntactic unit that appears in a sentence. In other words, the difference between a *śabda* and a *pada* is that of a 'word' and its 'fully inflected forms' [18]. A *śabda*/word is a language unit such as lexeme, word or word form whereas a *pada* is a word form that has inflected to mark its semantic relation with other words. Pāṇini categorizes a *pada* into two classes: 1) *subanta* and 2) *tiṅanta* [21, 9].

1. *Subanta*: A *subanta* class includes the participants of a sentence which inflect for marking various semantic relations. A *subanta* is formed by suffixation of nominal inflections called *sup*[1].
   The *subanta* class includes all nouns, pronouns, adjectives and adverbs etc. in it. Since, adverbs are indeclinables, they do not inflect for any case. That is why adverbs do not seem to carry any *sup* on surface. The sūtra (A.2.4.84) deletes the inflections attached to the adverbs. It suggests that at some point of time, adverbs also had been inflecting like other nominals. Hence, adverbs also fall in the *subanta* class.
2. *Tiṅanta*: This class includes words which mark some semantic relations and finiteness of the verb. In Sanskrit, verbs take *tiṅ*[2] suffixes to express tense, number, person, mood and voice. Similarly, in English, the auxiliaries and modal verbs when attached to a verb express tense, number, etc. hence correspond to *tiṅ*.

The primary objective of analyzing a sentence is to identify what role each part is playing. The role of the finite verb becomes important with relation to other participants. Therefore, Pāṇini takes only two classes.

Though this classification is mainly based on the surface realization of the words in Sanskrit, Pāṇini's concepts of grammar are not specific to Sanskrit. The concepts are rather generic and can be applied to other languages. The major theoretical concepts from PG would directly apply on other agglutinative languages, languages similar to Sanskrit. However, they can be extended to other languages as well.

In Sanskrit, the nominal inflections *sup* and the finite verb inflections *tiṅ* are realized through suffixation. However, the syntactic mechanisms for marking the semantic relations across words in a sentence might be different in different languages. Some languages such as Persian might have prepositions, some might have other kind of syntactic devices. For example, in Hindi, the relations of a noun to the verb or other nouns are marked through postpositions. From the Pāṇinian perspective, a *pada* in Hindi would be 'noun+postposition' [5].

[1] *Sup* is the acronym formed from the first and the last phoneme of the list of nominal suffixes.

[2] *Tiṅ* is the acronym formed from the first and the last phoneme of the list of verbal suffixes.

Sanskrit has a grammatical rule *apadaṁ na prayuñjīta* [11] which says: "a word which is not a *pada* should not be used in a sentence" [9]. Sobin's statement that "only phrases may be sentence fragments" [24] imposes a constraint similar to the statement *apadaṁ na prayuñjīta* for English. It is alluring to compare the two concepts *pada* of Sanskrit and 'phrase' of English.

Application of PG to other languages and finding out its effectiveness for machine translation (MT) is the task on hand. In this paper, we look at English from the Pāṇinian perspective. We investigate equivalent mechanisms of *sup, tiṅ* and *pada* in English phrases and compare them with the notions of *pada* and *samasta-pada* (compound) in Sanskrit.

We talk about the related work in Section 2. Section 3 describes the necessary conditions for *pada* formation according to PG. Section 4 defines *sup, tiṅ* for English and compares English phrases with the notions of *pada* and *samasta-pada* in Sanskrit. Section 5 shows a continuum between phrases and compounds. Section 6 shows how complex phrases are handled using the Pāṇinian perspective. Section 7 concludes the paper.

2 Related Work

Gangopadhyaya has analyzed noun phrases in Bengali and studied assignment of role and the kāraka theory following the Indian grammatical tradition. According to her: "The term phrase corresponds to the term *pada* in its minimal form but not in its expanded form, i.e. when a phrase is understood as a syntactic constituent consisting of more than one word." According to her, a single word phrase corresponds to a *pada* but a phrase that consists of more than one word does not correspond to a *pada* [12].

Gangopadhyaya does not account for phrases which consist of multiple words such as "brave soldiers", "very intelligent boy", etc.

According to Apte [2], the expression of a single idea is a word (*pada*) and the aggregation of two or more words without a subject or predicate is a phrase (*padasamuccaya*).

Apte calls a phrase as *padasamuccaya*, but does not give any linguistic account for it. He looks at Sanskrit from the English perspective, and therefore interprets a phrase as a group of multiple *padas*. But if we look at English from the Pāṇinian perspective, we find that a minimal/simple phrase corresponds to a *pada* or a *samasta-pada*. And, a complex phrase that is composed of two or more phrases corresponds to a *padasamuccaya*.

Local Word Grouping (LWG) [5] is a notion similar to *pada* found in literature. In LWG, word groups are formed on the basis of local (adjacent) word information for Indian languages.

Our assessment is based on flow of information where word groups are formed on the basis of neighboring syntactic inflections called *sup* and *tiṅ* for any language. This is grammatically more precise and also allows to find out syntactic elements that unite the words of a sentence into a meaningful unit.

To the best of our knowledge, we have not found any work that analyzes English phrases from the Pāṇinian perspective.

3 *Samartha* Theory of Pāṇini and its Relation to *Pada* and Phrase Formation

Pāṇini's *samartha* theory stands as a fundamental principle for any semantic and syntactic operation. According to Pāṇini, no grammatical operation can take place, be it *pada* formation or sentence formation, until and unless they qualify the condition of being *samartha* [21]. Thus the concept of *sāmarthya*[3] is a fundamental principle for any grammatical operation in a language string. The word *samartha* is used in the following two meanings:

1. *Ekārthībhāva sāmarthya*: It says that 'formation of a *pada* depends on unity of meaning' [26]. In this case, the *padas* having direct semantic relation become one *pada* as in compounds and primary and secondary derivatives. Here the word *samartha* means "organized together" (*saṁgatārthaṁ samarthaṁ*) and "fused together" (*saṃsṛṣṭārthaṁ samarthaṁ*) [17]. The objective of *ekārthībhāva sāmarthya* is to present compounds as one *pada* (*ekapada*) or one unit.
2. *Vyapekṣā sāmarthya* (meaning-interdependence): It says that "any operation pertaining to *padas* takes place if and only if the *padas* have direct semantic connection" [26]. In this case, *samartha* means "seen together" (*samprekṣitārthaḥ samarthaḥ*) and "bound together" (*sambaddhārthaḥ samarthaḥ*) [17]. For example, subject, verb, object etc. are seen bound together in a sentence. The objective of *vyapekṣā sāmarthya* is to show sentence as one unit. *Padas* seem to carry diverse meanings but a sentence indicates a single meaning.

For a word to stand in a syntactic structure, it is necessary to pass through one of these *sāmarthyas*.

Let us take the Sanskrit sentence (1) and examine how Pāṇini captures the flow of information through his grammar.

(1)	vīrāḥ	sainikāḥ	deśaṃ	rakṣanti
	brave.PL,NOM	soldier.PL,NOM	country.SG,ACC	protect.PR,3,PL

'Brave soldiers protect the country.'

In (1), the word *rakṣanti* is a *tiṅanta pada*. It is composed of the verbal base *rakṣ* and a *tiṅ* inflection namely -*anti*. A *tiṅ* inflection is assigned to a verb with respect to its compatibility (*sāmarthya*) with the doer/agent or theme/patient of the action. In the active voice, the *tiṅ* suffixes express the doer/agent of the action through agreement. The suffix -*anti* denotes active voice, third person, plural. When the doer is expressed by a *tiṅ* suffix, the sūtra (A. 2.3.46) [27] assigns nominative case (*prathamā vibhakti*) to the doer to express nominal stem

[3] The words *samartha* and *sāmarthya* are used interchangeably in Sanskrit grammar.

meaning (*prātipadikārtha*), gender (*liṅga*), or number (*vacana*) etc., of the doer. This also makes the nominal a *subanta pada*.

In (1), the words *sainikāḥ* and *vīrāḥ* are marked with nominative case (*prathamā vibhakti*) and plural number. The agreement between the verb *rakṣanti* and nominal *sainikāḥ* (soldiers) indicates that the *pada sainikāḥ* plays the role of the doer/agent of the action *rakṣanti* (protect).

Having the same *vibhakti*, the words *sainikāḥ* and *vīrāḥ* express the modified and modifier relation between them and also confirm semantic compatibility (*sāmarthya*) among them.

The theme/patient *deśaṃ* (country) is marked with the *sup, -am* (accusative case, singular number). It makes *deśaṃ* (country) a *subanta pada.*

From the above description, it is clear that a *pada* is a syntactic unit that takes a nominal or a verbal inflection called *vibhakti*, which explicitly marks the semantic relation of a word with another participant. In some cases, the *vibhakti* can also be NULL (zero) but it has to be present.

3.1 *Sāmarthya* in Phrases

While looking deeply at *samartha* theory and the concept of phrase in English, we noticed that both the theories capture the same aspect that is the coherence of words together in a significant sentence (*ekārthakatā*) but in slightly different ways. According to English grammar, a phrase is a sequence of words or sometimes a single word that functions as a single unit within a sentence [19, 16]. The words that are closely related to each other form a syntactic constituent.

According to *samartha* theory, for a word to become a *pada* or to form a word group with other word/s such as in compounds, it has to have direct semantic relation with the other words in the sentence. Thus, the notion of *sāmarthya* is the linguistic driving force behind formation of both phrases and the *padas.* The rest of the paper explores it in detail.

This notion of *pada*, though developed for Sanskrit which has a rich inflectional and derivational morphology, can be applied to any language. If we apply it to English which is morphologically not so rich, we have to analyse English sentences from the Pāṇinian perspective, especially in terms of *sup*, *tiṅ*, and *pada.*

4 *Subanta Pada* and *Tiṅanta Pada* in English

In Sanskrit, *sup* carries information of number and the case marker. However, English has different sets of morphemes for marking number and case information. In English, the number information is marked through a suffix '-s' for plural and '-0' (NULL) for singular and the case is realized through prepositions or through 'generalized *vibhakti*' [3, 4] in terms of the position of the subject or object [1]. For example, in "to the boy" and "to the boys", the preposition "to" marks case information and the inflection "-s" marks number information.

In English, the *tiṅ* inflections are realized through auxiliaries and modal verbs. The *tiṅ* suffixes also express the role of one of its participants through agreement.

Attachment of a nominal or a verbal inflection to a nominal or verbal entity makes it a *subanta* or a *tiṅanta pada* respectively.

Let us take the English sentences in (2) and analyse it using Pāṇinian primitives such as *sup*, *tiṅ* and *pada*.

(2) She gave books to Mohan.

Figure 1 shows the constituency tree diagram for sentence (2). The prepositional phrase "to Mohan" carries preposition "to" as a *sup*, hence, it corresponds to a *subanta pada*. The verb *gave* has *-ed* as a *tiṅ* suffix, hence it can be considered a *tiṅanta pada*.

The words "She" and "books" are the subject and object respectively and do not seem to have any explicit case marker. Bharati et al (1996,1998) have shown that English has the notion of 'generalized vibhakti' which corresponds to the *sup* suffixes in Sanskrit. The 'generalized vibhakti' is realized either through subject[4] or object positions or through prepositions. Thus in sentence (2), *She* occurring at the subject position seems to carry no *sup*, but according to Bharati et al. (1996,1998), since it occurs at the subject position, it carries a generalized vibhakti in terms of subject position, hence, it is a *subanta pada*. Similarly, the object "books" carries a generalized vibhakti in terms of object position, hence, it is also a *subanta pada*. See Figure 1, where each box represents an independent *pada*.

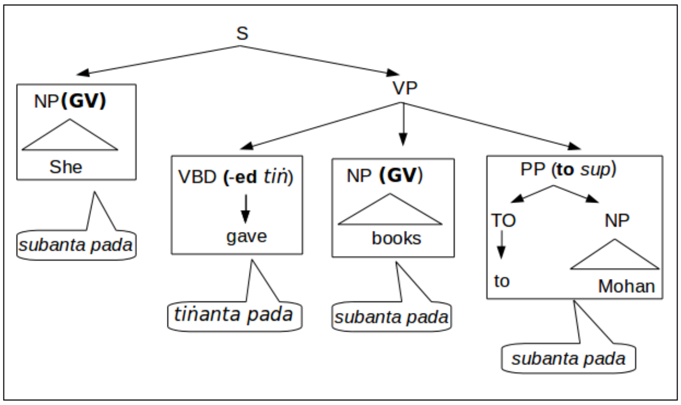

Fig. 1. *Pada* information in tree diagram for (2). In this figure GV stands for 'generalized vibhakti'

[4] In linguistics, the notion of subject in ILs is much debatable [7].

In sentence (2) each phrase consists of a single lexical item, therefore, we can say that a phrase with a single lexical item corresponds to a *pada* from the Pāṇinian perspective. But a phrase is not always a single word.

A phrase is defined as a word or a sequence of words that functions as a single unit within a clause/sentence [16, 19]. Given this definition, how would PG handle a phrase that consists of more than one word? For instance,

(3) Four things scientists have been exploring about the incredibly awesome moon.

In sentence (3), the word group "four things", and "scientists" are NPs; "have been exploring about the incredibly awesome moon" is a VP which consists of a finite verb group, and a prepositional phrase. From the above description, one can say that the NP, "scientists" being at subject position carries a *sup* in term of subject position. So, "scientists" is treated as a *pada*. The verb group "have been exploring" contains "explore" as a verb and "have_been_ing" as a *tiṅ*. So, "have been exploring" is treated as a *tiṅanta pada*. But what about the NP "four things" which consists of more than one word and does not have any preposition to mark it a *pada*? And, what about the prepositional phrase "about the incredibly awesome moon" which though has a preposition/*sup* but at the same time contains one more phrase in it, the adjectival phrase (ADJP) "incredibly awesome"? How many *padas* should we consider in the constituents "four things" and "about the incredibly awesome moon"? In order to resolve such issues, let us first look at the characteristics of Sanskrit compounds (*samasta-pada*).

4.1 Four Characteristic Features of the Compounds (*samasta-pada*)

A Sanskrit compound has following four characteristic properties:

1. *Sublopa* (elision of internal *sup/vibhakti*): Elision of internal *sup* as opposed to an *asamasta-pada* (sentence) takes place in a *samasta-pada*. An uncompounded word group is also called vākya (sentence) in Sanskrit. Only the final element receives case inflection in compounds. For example, the genitive case marker (*ṣaṣṭhī-vibhakti*) is deleted in the samasta-pada *rājapuruṣaḥ* (king-man), whereas in the uncompounded word group *rājñaḥ puruṣaḥ* (king's man), the genitive case marker is not deleted.
2. *Avyavadhāna* (no intervention by other word (*pada*)): Intervention by any other word (*pada*) does not take place in a *samasta-pada* but in an uncompounded word group it can take place. For example, one can say *rājñaḥ ṛddhasya puruṣaḥ* (man of a rich king), where *rājñaḥ puruṣaḥ* has been intervened by *ṛddhasya* a modifier of *rājñaḥ*, but in a compound, *ṛddhasya* cannot modify *rājñaḥ*. One cannot say **rāja-ṛddhasya-puruṣaḥ*.
3. *Niyatapaurvāparya* (fixed word order): The words in a *samāsa* occur in a fixed order. But in a sentence the *padas* can occur freely; *rājñaḥ puruṣaḥ* or *puruṣo rājñaḥ*

4. *Aikasvarya* (accent/stress): All the words in a *samasta-pada* have only one accent/stress. For example, in *rājapuruṣaḥ* the stress is on the final vowel/syllable, but in a sentence, *rājñaḥ puruṣaḥ*, both the words are stressed independently.

Having looked at the characteristic properties of *samasta-pada* (compounds), let us now look at the English phrases and see whether they are comparable with *samasta-pada* or not. Most phrases have all the four properties listed above. We will examine them one by one. For example,

No internal preposition/*sup* is present between the words in the phrase "four things". Thus, the absence of a *vibhakti* suggests that the internal preposition must have been deleted. So, there is a *sublopa*.

No other phrase can occur within a phrase, i.e. we cannot say "*four to explore things". So, the phrase has *avyavadhāna* feature of compounds.

The words have fixed order. One cannot say "things four". So, they follow the principle of *niyatapaurvāparya*.

In English, stress is on the first word in common phrases and on the noun in descriptive phrases. Table 1 shows stress variations in common and descriptive phrases. The words in bold have stress in these phrases[5].

Common phrase	Descriptive phrase
a **sports** car	a small **car**

Table 1. Showing stress variations for common and descriptive phrases.

In our example, in the phrase "four things", the stress is on **'things'**, 'four **things**'. So, there is *aikasvarya*.

Being able to see the similarities between a phrase and *samāsa/samasta-pada* (compound) only tells us that the components of a phrase have some semantic relation among them. But, even for a compound to participate in a sentence, it has to become a *pada*. It should have some *vibhakti* to express its relation with other participants of the sentence.

It seems that the phrase "four things" neither has any overt *vibhakti* nor any generalized *vibhakti*. Then, how would it pass the test of being a *pada*?

In (3), the verb "explore" has two arguments. One argument is represented by the subject "scientists". But the second argument represented by "four things" is not at the object position. It has moved leftward to the initial position of the sentence for topicalization [16]. The NP "four things" originated at the object position of "explore" leaves a trace at object position which encodes the relation of the moved element. To put it in other words, the topic position also assigns *vibhakti* to the topicalized constituent. Thus, the NP "four things" has its generalized *vibhakti* in terms of topic position and that is why it is a *pada*.

In the PP "about the incredibly awesome moon", it is not the case that an external element/phrase "incredibly awesome" has intervened between "the"

[5] See "Learn English with Speak Method, URL: www.speakmethod.com/syllablestresssetphrases.html"

and "moon" hence, there is a violation of one of the compounding characteristics called *avyavadhāna*. Rather, it is an example of embedded compounding. Here, first the words "incredibly" and "awesome" form a compound and then the compound "incredibly awesome" forms another compound with the words "the" and "moon". After that, the compound "the incredibly awesome moon" takes the preposition "about" as a *sup* which makes it a *pada*.

To summarize, we can say that the constituent "scientists" in sentence (3) occurs at the subject position, hence, as stated by Bharati et al. (1996,1998), it carries a 'generalized vibhakti' in terms of subject position. The NP "four things" gets its *sup* inflection from its trace at object position. Also, if the constituent "four things" occurs in a prepositional phrase, both the words take only a single preposition as in *I gave you money* ***for four things***. Therefore, the whole group, "four things" will be treated as a single *subanta pada*. The phrase "about the incredibly awesome moon" can also be taken as a *subanta pada* which carries the preposition "about" as a *sup*.

From the above observations, we can say that complex English phrases come close to compound constructions in Sanskrit, except that in Sanskrit, a compound becomes a single word whereas in English, the phrasal components maintain multiple word status. This hardly matters. Because, compound constructions in English cover a whole range of written styles such as one word as in "milkman" (man who delivers milk), hyphenated as in "milk-fever" (disease caused by lack of the calcium contained in milk) and with white space as in "milk bottle" (bottle for containing milk). Hence unlike Sanskrit, *ekapadībhāva* [20] 'becoming one word' of more than one words is not a characteristic property of English compounds. In fact in [14] and [13], Giegerich argues that the distinction between compound and phrase is neither necessary nor possible in English. Therefore, simple English phrases can be treated as *padas*.

So, we can say that English phrases share the properties of a compound to some extent. However, unlike Sanskrit compounds, they maintain more than one word status. Thus, they are not compounds in the strictest sense. Therefore, we call them 'quasi-compounds' (*ardhasamāsa*).

We have already seen that in (1), the modifier and modified relation between *vīrāḥ* (brave) and *sainik=aḥ* (soldiers) is expressed by attachment of the same case marker to both the words. Similarly, if the head *sainika* takes some other case marker to express its relation, the modifier *vīra* also takes the same case marker such as *vīrān sainikān* (to brave soldiers), *vīrebhyaḥ sainikebhyaḥ* (for brave soldiers), *vīrāṇām sainikānām* (of brave soldiers) etc.,. On the other hand, English attaches a single preposition to all the members of the constituent "brave soldiers": "by brave soldiers", "for brave soldiers", "of brave soldiers". As a result the positions of the constituent members get fixed and phrases are formed.

As explained in Section 1, a *tiṅanta pada* is formed by adding *tiṅ* inflections to the verbal bases.

As mentioned in Section 4.1, intervention of any external word/*pada* is not allowed in a *samasta-pada* but the finite verb groups are exceptions to this condition, where an adverbial phrase can intervene in a finite verb group. For example, take the expression in (4):

(4) 'have been slightly changed'

In (4), the adverb "slightly" is embedded in the *tiṅanta pada*, "have been changed". But then, since a *tiṅanta pada* is not a compound, other words can intervene.

It should be noted that except adverbs, no other words can intervene in a *tiṅanta pada.* That is why, the expressions like: "*have **to operate** been changed", "*have been **to operate** changed", '*have been **to Paul** changed", etc., become ungrammatical.

Identification of *tiṅanta padas* helps in forming a verb group whereby translation of verb and its suffixes can be handled properly

5 Spectrum of Flexibility in Compounds

If we look at the various types of expressions, there appears to be a continuum from 'flexible' expressions to completely 'fixed' expressions in compounds. For example, in Sanskrit, *januṣāndhaḥ* (blind from birth, born blind), *alpānmuktaḥ* (loosed from a little distance), *vācoyuktiḥ* (appropriate speech) etc., are some of the expressions where there is no deletion (*lopa*) of the internal *sup vibhaktis* and the compound meaning is similar to that of the canonical phrasal paraphrase (*vigraha-vākya*) [15].

Presence of a *vibhakti* expresses the relation between words. In compound expressions, internal *vibhaktis* and number information are not so important [2, 25, 26]. The relations among members of a compound are expressed through positions of the words. That is why *vibhaktis* can vanish. Even in expressions like *stokānmuktaḥ, goṣucaraḥ* etc., where the internal *vibhakti* is not deleted, we use the same expression *stokānmuktaḥ*, even if we wish to say *stokābhāṃ muktaḥ* (dual) or *stokebhaḥ muktaḥ* (plural) [26].

In *rājapuruṣaḥ* etc., even after deletion of the internal *sup*, the potency of the *sup* takes place by *pratyayalakṣaṇa* (A.1.1.62), that is, the operations pertaining to a suffix take place even after deletion of the suffix [27]. Therefore, the component *rājan* is treated as a *pada* for morphological operations, as a result elision of the letter *n* by (A. 8.2.7) can be seen on the surface. Expressions like *brāhmaṇakambalaḥ, yūpadāru* etc., also undergo similar operations but no effect of the internal *sup* can be structurally seen on the surface. In such expressions, the compound meaning is similar to their components and one can construct synonymous compounds using synonyms. For instance, *rājabhṛtyaḥ* is the paraphrase of *rājapuruṣaḥ* and *dvijakambalaḥ* is the paraphrase of *brāhmaṇakambalaḥ.*

On the other hand, there are expressions like *kṛṣṇasarpaḥ* which are completely inflexible. These expressions have all the characteristics of compounds but irrespective of the meanings of the words involved in compounding the expressions give a more specialized meaning. In our example, the compounding words *kṛṣṇa* (black) and *sarpa* (snake) leave their meanings and give the special meaning "cobra". Paraphrasing is also not possible in such cases. Such compounds are called *nityasamāsa* (completely fixed expressions). Table 2 gives an overview of spectrum in Sanskrit compounds. It illustrates the four afore mentioned compound properties plus 'one word', 'multiple word' status and 'paraphrasing' with examples where, you will notice that as one moves from left to right the degree of flexibility of the expressions varies.

Table 2. Showing flexibility spectrum in Sanskrit expressions

	aluksamāsa	*samāsa*	*nityasamāsa*
	vācoyuktiḥ	*rājapuruṣaḥ*	*kṛṣṇasarpaḥ*
sublopa	x	✓	✓
avyavadhāna	✓	✓	✓
niyatapaurvāparya	✓	✓	✓
aikasvarya	✓	✓	✓
one word	✓	✓	✓
multiple words	x	x	x
paraphrasing	✓	✓	x

5.1 Flexibility Spectrum in English Compounds

We claim that all languages including English have flexibility spectrum in compounding. In English, the level of variation is similar to Sanskrit or even higher. We have already seen that English phrases fall under 'quasi-compound' class. The 'quasi-compounds' show the highest degree of flexibility in English.

Instances of *aluksamāsa* are also found in English. For example, in "kinsman" (a blood relative, especially a male), marksman" (a person who is skilled in shooting at a mark) etc., the compound members "kins" and "marks" are possessive forms "kin's" and "mark's" [10]. The possessive suffix "-'s" does not vanish in compound forms.

The compounds like "blackbird" denoting a bird of a particular species [10] fall under *nityasamāsa* class because such compounds give a very specialized meaning different from the compound members.

Table 3 shows the flexibility spectrum in English compounds.

6 Handling Complex Phrases from The Pāṇinian View Point

After finding evidence for simple phrases, it is time to move towards more complex phrases such as "the symbol of mature wisdom", where the phrase is composed of two different phrases: an NP "the symbol" and a PP "of mature wisdom". Or other complex phrases such as "the boy who came from Delhi yesterday", where the noun phrase is composed of an NP and a relative clause. The

Table 3. Showing flexibility spectrum in English expressions

	ardhasamŝa	*aluksamāsa*	*samāsa*	*nityasamāsa*
	a good boy	kinsman, marksman	lawn tennis, bird-cage, football	blackbird
sublopa	✓	x	✓	✓
avyavadhāna	✓	✓	✓	✓
niyatapaurvāparya	✓	✓	✓	✓
aikasvarya	✓	✓	✓	✓
one word	x	✓	✓	✓
multiple word	✓	x	✓	x
paraphrasing	✓	✓	✓	x

clause in its turn has multiple phrases in it. As shown above, all the member phrases of a complex phrase fall under the *subanta* or *tiṅanta pada* class.

The PP “of mature wisdom” has a direct semantic connection with the NP “the symbol”.

The minimal phrases forming a complex phrase are seen bound together. That is why the tests like movement and substitutions etc. apply to the entire complex phrase and as a result the whole phrase moves bound together or is substituted as a whole and represents a single meaning which we call *ekārtha*. The fact that the entire complex phrase comes under the scope of a single preposition (*sup*) suggests that the entire complex phrase represents one unit.

Since a complex phrase is composed of two or more phrases/*padas* and all the *padas* maintain their *padaness* i.e. the *subantaness* (*subantatva*), one can not call that complex phrase a pada. The name we've given to such phrases is *subantamukhyaviśeṣyaka-ekārthaka-padasamuccaya.* It means a group of phrases (*padasamuccaya*) which has a single meaning (*ekārthaka*) where a *subanta pada* is the head of that particular complex phrase (*subantamukhyaviśeṣyaka*). The hierarchic organization (internal phrasal structure) within a sentence also supports the *subantamukhyaviśeṣyaka-ekārthaka-padasamuccaya* view.

The concept of *vyapekṣā sāmarthya* in Pāṇini is a general principle to capture meaning interdependence in a sentence. It not only captures the meaning interdependence among nominals but also connects the *subanta padas* with the *tiṅanta pada.* This generality of *vyapekṣā sāmarthya* might seem to be overcovering the VPs also but the element *subantamukhyaviśeṣyaka* in the term *subantamukhyaviśeṣyaka-ekārthaka-padasamuccaya* restricts it only to the *subanta padas.*

Identification of *padas* helps in the demarcation of syntactic units in a sentence. Once the syntactic units are identified, source language to target language generation becomes easy, especially, when target language is morphologically richer than the source language, such as Hindi and English. Similar to English, Hindi also has ‘quasi-compounds’ phenomenon. For instance, let us take the expressions in (5).

(5) a. acchā bacca̅
good.SG,DIR child.SG,DIR
'good child'

b. acche bacce ke liye
good.SG,OBL child.SG,OBL for
'for the good child'

In (5-a), neither of the words has any case marker, therefore, both the words are in direct case [5]. But in (5-b), the head *baccā* (child) has the case marker *ke liye* attached to it. So, the head *baccā* (child) is in oblique case. Since no covert case marker is present after the adjective *acche*, it should be in direct case. But that is not true in this case. But if we consider it to be a 'quasi-compound', we can say that *sublopa* (elision of internal case marker/*sup*) has taken place in this expression. And, by the principle of *pratyayalakṣaṇa*, the operations pertaining to *sup* takes place. Hence, the modifier *acche* is in the oblique case.

Identification of syntactic units is also helpful in reordering partially free word order language [9].

7 Conclusions

We have analyzed English phrases based on the Pāṇinian perspective. We have defined nominal inflection *sup*, and finite verb inflections *tiṅ* for English and compared English phrases with the notion of *pada* and *samasta-pada* in Sanskrit. We have shown that a single word phrase directly corresponds to the concept of *pada* in Sanskrit and a complex English phrase (a phrase consisting of more than one word) corresponds to compounds. The study shows that the insights from Pāṇinian Grammar can be used to analyze any language from information theoretic point of view. We have also briefly shown its effectiveness in machine translation.

Acknowledgments. We are immensely grateful to Prof. Vineet Chaitanya for sharing his insights and for his continuous guidance during this research. We would also like to thank Prof. Amba P Kulkarni for the discussions in the early stage of the research, to Sirisha Manju, Roja Laxmi and Mahalaxmi for their assistance in implementation of the concepts for English-Hindi machine translation, to Pratibha Rani for her comments on an earlier version of the manuscript, and to Prof. Aditi Mukherjee and Prof. Shatrunjay Rawat for their feedback on the final version of the paper. However, we have not yet fully incorporated their comments into this version of the paper, we wish to incorporate them in future versions of the paper.

References

1. Anantpur, A.P.: Anusaaraka: An approach for MT taking insights from the Indian Grammatical Tradition. Ph.D. thesis, University of Hyderabad (2009)

2. Apte, V.S.: The student's guide to Sanskrit composition. Chowkhamba Sanskrit Series Office (1963)
3. Bharati, A., Bhatia, M., Chaitanya, V., Sangal, R.: Paninian grammar framework applied to English. Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur (1996)
4. Bharati, A., Bhatia, M., Chaitanya, V., Sangal, R.: Paninian grammar framework applied to english. South Asian Language Review 8(1), 1–23 (1998)
5. Bharati, A., Chaitanya, V., Sangal, R.: Natural language processing: a Paninian perspective. Prentice-Hall of India New Delhi (1995)
6. Bharati, A., Kulkarni, A.: Information coding in a language: Some insights from pāṇinian grammar. Dhiimahi, Journal of Chinmaya International Foundation Shodha Sansthan I(1), 77–91 (2010)
7. Bharati, A., Kulkarni, A.: 'subject' in english is *abhihita* (2011)
8. Bharati, A., Sukhada, Sharma, D.M., Paul, S.: Sanskrit and Computational Linguistics, chap. Anusāraka Dependency Schema from Pāṇinian Perspective. D. K. Publishers (2015)
9. Bharati, Akshar, S.J.P.P.S., Sharma, D.M.: Applying sanskrit concepts for reordering in mt. In: Proceedings of the ICON2015 (2015)
10. Bloomfield, L.: Language. Motilal Banarasidass Publishers Private Limited, Delhi (1994)
11. Dvivedi, K.: Rachanānuvādakaumudī. Vishwavidyalaya Prakashan (1953)
12. Gangopadhyay, M.: The noun phrase in Bengali: assignment of role and the kāraka theory. Motilal Banarsidass Publishers. (1990)
13. Giegerich, H.: How robust is the compound-phrase distinction? stress evidence from bi-and tripartite constructions in english. Linguistics 2, 65–86 (2008)
14. Giegerich, H.J.: Attribution in English and the distinction between phrases and compounds (2006)
15. Gillon, B.S.: Exocentric (bahuvrīhi) compounds in Classical Sanskrit. In: Proceedings, First International Symposium on Sanskrit Computational Linguistics. pp. 1–12 (2007)
16. Haegeman, L., Guéron, J.: English grammar: A generative perspective. Blackwell Oxford, England (1999)
17. Joshi, S.: Patanjali's vyakarana-mahabhasya. Samarthahnika (1968)
18. Kapoor, K.: Dimensions of Pāṇini Grammar: The Indian Grammatical System. DK Printworld (2005)
19. Kroeger, P.R.: Analyzing grammar: An introduction. Cambridge University Press (2005)
20. Mahavir: Pāṇini as Grammarian: With Special Reference to Compound Formations. Bharatiya Vidya Prakashan (1978)
21. Mahavir: Samartha Theory of Pāṇini and Sentence Derivation. Munshiram Manoharlal Publishers (1984)
22. Sharma, R.N.: The Aṣṭādhyāyī of Pāṇini: Introduction to the Aṣṭādhyāyī as a Grammatical Device, vol. 1. Munshilal Manoharlal Publishers (1987)
23. Singh, J.D.: Pāṇini, his description of Sanskrit: An analytical study of Aṣṭādhyāyī. Munshiram Manoharlal Publishers (1991)
24. Sobin, N.: Syntactic analysis: the basics. John Wiley & Sons (2010)
25. Speijer, J.S.: Sanskrit Syntax. Motilal Banarsidass (1886)
26. Varma, S.: Vyākaraṇa Kī Dārśanika Bhūmikā. Munshiram Manoharlal, New Delhi (1971)
27. Vasu, S.C.: The Aṣṭādhyāyī of Pāṇini. Motilal Banarsidass Publishers (1996)