Chapter 7

Conclusions and Future Work

The research described in this document was set out to explore Pāṇinian perspective to information dynamics in language. Its main emphasis was on mapping structures between English and Hindi and finding out the way languages encode information. It focused on application of insights from the Pāṇinian Grammar for Natural Language Processing for faithful transference of information in source language text to target language in machine translation.

We talked about the following two types of *padas*, primary syntactic units in Sanskrit:

1. *Prātipadika* + *sup* = *subanta pada*
   (nominal stem + nominal inflection = nominal *pada*)

2. *Dhātu* + *tiṅ* = *tiṅanta pada*
   (verbal stem + verbal inflection = verbal *pada*)

We saw that the notion of *pada* gives a lot of power to encode the grammatical relationships among the words in a sentence and flexibility for word ordering.

Taking the insights from the concept of *pada*, we talked about the syntactic mechanisms that natural languages use to encode the grammatical relationships among the words in a sentence. This includes:

- **Overt Mechanisms:** Overt mechanisms are those where languages use explicit morphemes to express the relations among the words, such as (a) derivational morphology and (b) inflectional morphology.

- **Covert Mechanisms:** Sometimes languages use implicit mechanisms like relative positions of the words to encode the relationships among the words. In 'covert mechanisms', languages fix or freeze the positions of the words with respect to other words. This mechanism can be further classified into two classes:

  1. Implicitness when language lacks explicit morphemes to express a linguistic phenomenon, Or
  2. Implicitness to bring brevity/economy to the expressions

Languages gain brevity in this process but at the same time loose the freedom of word ordering.

In 'overt mechanisms', we focused mainly on inflectional morphology for syntactic analysis of the sentences. We identified the nominal and verbal inflections for English based on the Pāṇinian primitives *sup* and *tiṅ* to identify the *padas*, the primary syntactic units in English. We found that besides the overt morphemes like prepositions or case endings, English also includes covert morphemes like relative positions of the participants in a sentence. In English, following positions play an important role as implicit morphemes: 'subject position', 'object position' and 'topic position'. Both, the explicit as well as the implicit mechanisms like relative positions are termed 'generalized vibhakti'. Thus, a *pada* in English is:

1. 'Generalized vibhakti' + nominal stem = *subanta pada*
2. (a) Verb + verbal inflection = *tiṅanta pada* OR
   (b) Auxiliary + verb = *tiṅanta pada* OR
   (c) Auxiliary + verb + verbal inflection = *tiṅanta pada*

We claimed that single word English phrases (simple English phrases) correspond to the notion of *pada* and complex phrases, phrases that consist of more than one word, corresponds to compound constructions, the *samasta-padas*. We verified this claim by examining multiple word phrases by application of the four characteristic properties of the compounds (*samasta-padas*) viz.

1. *Sublopa* (Elision of Internal *Sup*/*Vibhakti*)
2. *Avyavadhāna* (No Intervention by External Word/*Pada*)
3. *Niyatapaurvāparya* (Fixed Word Order)
4. *Aikasvarya* (One Accentuation/Stress)

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

We discussed about the continuum in terms of structural and semantic realization of the compound expressions. Structurally, compounds range from expressions having internal *vibhaktis* to expressions where internal *vibhakti* is elided. Semantically, they range from compounds that are fully compositional in meaning to the compounds that are non-compositional.

We developed the Anusāraka Dependency Schema. The research behind developing this schema was motivated for a common representation of the dependency parse outputs for English to Indian languages, especially English to Hindi MT system Anusāraka. Various translation sub-tasks such as Gender Number Person agreement, Word Sense Disambiguation, reordering of the source language sentence according to the target language sentence, etc., require dependency parse labels as input especially in a

rule based system. However, different parsers have different output schemes which differ in relation labels as well the number of the dependency labels. Bringing parsers' output into one uniform notation provides the system an ability to plug-in any of these parsers without modifying the other subsequent translation modules, thereby, avoiding large amount of manual work which is costly as well as time consuming.

We talked about the transfer grammar approach for English-Hindi word ordering and generation. We talked about the structural divergences between English and Hindi and deliberated a three stage reordering approach that takes mainly the constituency parse and *padas* as input at first stage and manipulates them to handle the major structural divergences between the two languages. At this stage, we mostly do mirroring of some of the English phrases, such as reversal of the verb phrase to handle the differences like head-initial and head-final nature of these languages. At second stage, we take help of dependency parse as well to re-arrange the structures that tend to precede or follow some other expressions. For instance, the destination phrase tends to occur immediately before the verb in Hindi as shown in the following example:

(92) He **took** her **home** from the party.
vaha **le āyā** use **ghara** se pārtī
'*vaha pārtī se **ghara use le gayā**.'

'vaha use pārtī se **ghara le gayā**.'

Insertion of some of the words like relative pronouns and subordinating conjunctions also takes place at stage two. For example, in the following sentence, relative pronoun *jise* (who) is inserted in Hindi translation which is not present in the source language English.

(93) Rama saw the man you love.
Rama.NOM dekha.PT vaha ādamī.SG āpa prema kara.PR
*Rama ne vaha ādamī dekhā āpa prema karate haiṃ.

'Rama ne vaha ādamī dekhā **jise** āpa prema karate haiṃ.'

After lexical substitutions according to the target language, if required, the sentence goes to the third stage where splitting or movement is done for fluency enhancement of the words in target language.

We have also shown the usefulness of the notion of *pada* and *samasta-pada* in target language word generation, especially, how it helps in generating the appropriate grammatical cases: the direct and oblique cases for Hindi.

We provided an overview of the tools built during this research. These tools analyse the source language sentences as well as generate target language. Its center of attention were the tools that accomplish following major linguistic tasks:

- A brief discussion on *pada* formation in Sanskrit
- Analysis of English phrases from the Pāṇinian point of view

- Identification of overt and covert mechanisms that languages use to encode the semantic information through syntax
- Identification of primary syntactic units called *padas* that can frame sentences.
- Identification of continuum in terms of structural and semantic realization of the compound forms
- Development of a standard dependency schema for mapping English parsers' output for robustness. Translation modules like Gender, Number, Person agreement, Word Sense Disambiguation, reordering of the source language sentence according to the target language sentence, etc. take dependency parse as input.
- Development of an English-Hindi word ordering tool using the concept of *pada* and reordering rules based on "mirror structure" along with some exceptional rules that arrange the *padas* into a fluent Hindi word order generation.
- Presentation of experiments, evaluation process and results for overall system comprehensibility test.
- Describing the experiments done for reordering rule accuracy for English-Hindi language pair with comparison to statistical system Google Translate.

The experiments obtained better results than the SMT system Google Translate which show that the rules are accurate enough to enhance the translation quality.

The main focus of this research is on English-Hindi language pair. But we claim that the research carried out in this thesis is generic enough to be applied to any English-Indian language pair. This is because most of these languages have many common features and the transparency of the tools built during this research allows to cover diverse cases easily. We have shown this through an English-Telugu machine translation sample.

In future, we plan to adapt these tools for English-Indian languages such as English-Marathi, Punjabi etc. We also plan to automate the dependency mapping tool, so that any new dependency parser can be mapped into Anusāraka dependency schema in less time and effort. Also, some work can be done in combining the strengths of various parsers to obtain more benefit.