Construction Grammar approach for Tamil dependency parsing

Thesis submitted in partial fulfillment
of the requirements for the degree of

MS by Research
in
Computational Linguistics

by

Vigneshwaran M
201350872
vigneshwaran.m@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
OCTOBER 2016

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Construction Grammar approach for Tamil dependency parsing" by Vigneshwaran M, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. Dipti Misra Sharma

To *Amma* and *Appa*

Acknowledgments

It is quite fascinating to wonder how mere structural arrangement of units such as sounds, words etc. are capable of representing abstract, meaningful ideas. Equally fascinating is the idea that with sufficiently good understanding and representation, this mechanism could be taught to machines as well. I would like to thank my professors in IIIT who introduced me to Linguistics and NLP - Professor Dipti Misra Sharma, Dr. Radhika Mamidi, Dr. Kishore Prahallad and Dr. Soma Paul for the learning I had from their courses.

Thanks a lot to my friends - Nikhilesh, Ganesh, Arpit, Vandan, Pruthwik, Litton, Devadath, Maaz, Silpa - for all the good times we had together. I am extremely grateful to Ganesh Katrapati and Devadath V for the many discussions, brainstorming sessions and presentations we did together. I cherish my experience with Ganesh for his constant companionship and collaboration when we explored Construction Grammar for Tamil and Telugu syntax. He was open to new ideas, patient when my initial ideas were vague at their best, offered his insights, helped organize the ideas better and formulate them more coherently. I thank Tejaswinee Kelkar for reading the earlier manuscripts and providing her feedback and suggestions for improvement; Himanshu Sharma, Nikhilesh Bhatnagar, Pruthwik Mishra who helped me during various stages of research; Litton J Kurisinkel with whom I had fun working on local discourse units for multi-document summarization; Ratish who provided valuable feedback with paper reviews. I am thankful to all these amazing people for what they are.

And most importantly I thank my guide Professor Dipti Misra Sharma for her constant support, for her encouragement, for her disagreements during our many presentations and discussions, critical reviews and feedbacks without which the current work would not have been possible. I also am thankful to Dr. Manish Shrivastava, Dr. Radhika Mamidi and Dr. Soma Paul for their inputs during research progress presentations. Of course, above all, I am indebted my mother, father and brother without whose understanding and support none of my life choices would be possible.

Abstract

Syntactic parsing in NLP is the task of working out the grammatical structure of sentences. Some of the purely formal approaches to parsing such as phrase structure grammar, dependency grammar have been successfully employed for a variety of languages. While phrase structure based constituent analysis is possible for fixed order languages such as English, dependency analysis between the grammatical units have been suitable for many free word order languages such as Indian languages. All these parsing approaches rely on identifying the linguistic units based on their formal syntactic properties and establishing the relationships between such units in the form of a tree.

Dravidian languages which are spoken in Southern India are morphologically-rich, agglutinative languages whose characterization on purely structural terms such as adjectives, adverbs, conjunctions, postpositions as well as traditional interpretations of tense and finiteness pose problems in their syntactic analysis which are well-discussed in literature. We propose that the morpho-syntactic structures of Dravidian languages are better analysed from the theoretical perspectives of "Cognitive Grammar" or "Construction Grammar" where every grammatical structure is treated as a symbol that directly maps to meaningful conceptualizations. In other words, natural language is not treated as a formal system but as a functional system that is entirely symbolic or semiotic right from lexicon to grammar.

Through linguistic evidences we point out that morpho-syntactic structures in Dravidian languages have their basis in meaningful discourse conceptualizations. Subsequently we hierarchically arrange all these conceptualizations into *construction schemas* that exhibit multiple-inheritance relationships and we explain all concrete morpho-syntactic structures as instances of these schemas. Based on this fresh theoretical grounding, we model parsing as automatic identification of meaningful dependency relations between such meaningful *construction units*. We formulated an annotation scheme for labelling the construction units and dependency relations that can exist between these units. Our approach to full parser annotation shows an average MALT LAS of 82.21% on Tamil gold annotated corpus of 935 sentences in a five-fold validation experiment. We conducted experiments by varying training data size, annotation scheme, length of a sentence in terms of number of chunks, granularity of tags and report the parser results of these scenarios. Finally, we build a pipeline with splitter, construction labeller, grouper as intermediate layers before MALT parser input and release the working full parser module.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A common theoretical assumption in dependency parser implementations is that grammar is a formal system of rules which arranges the morpho-syntactic units that make up a sentence. In post-Chomskyan structuralist tradition, syntax is understood as an autonomous formal component in the language architecture with its own rules of arrangement[15]. We see that in generative approach to language, abstract forms of linguistic *structure* are represented in a formal meta-language with grammar formulated as a self-contained study independent of semantics.

We observed that Dravidian languages, spoken in South India, exhibit interesting morphological properties which can be better characterized from a functional approach to syntax such as Construction Grammar[1], [2] or Cognitive Grammar[3]. According to the theoretical assumptions of Construction Grammar, every grammatical formative in a language is analysed as a meaningful mapping between its form and function. We identified that the set of morphological inflections that occur in Dravidian languages can be mapped to a set of meaningful 'construals' i.e. as a form-function pairing. We will introduce and talk about these 'construals' in chapter 3. Wherever morphs can be grouped together as one meaningful construal let us call that as one construction unit. Our idea is that to parse a sentence, one has to identify which morphs should be grouped together as one meaningful construction unit, identify the appropriate construction label for the grouped unit, chunk these construction units and finally identify the dependency relations between these chunks. At the outset, the aim of the thesis is threefold:

- to outline some of the learning issues that we encounter in our current formal approaches to parsing Dravidian languages
- to propose an alternative theoretical approach with relevant linguistic evidences
- to apply these theoretical insights on full parser annotation scheme and verify if the same problematic syntactic scenarios are handled better in the new approach

It should be remembered that we are not changing the machinery used for parsing i.e. the same MALT parser is used to learn the dependency relations, but we are just changing the theoretical framework used to analyse and represent the dependency structure of a sentence.

The thesis is organized as follows. Chapter 2 describes the various approaches to parsing and proceeds to point out the linguistic constructions that pose difficulties in processing Tamil. We illustrate through examples that there are many learning issues and problematic cases which are theoretical in nature and suggest that an alternative theoretical perspective based on discourse conceptualization is needed.

Chapter 3 introduces the theoretical notions of functional approaches to grammar, in particular *Cognitive Grammar* and *Construction Grammar* from which we got our inspiration for a purely functional approach to Dravidian syntax. In this chapter, we discuss the basic construction schemas that are fundamental and most abstract in nature.

Chapter 4 expands the basic schemas and derives more subtypes of construction schemas from the generic ones. We propose a total of 29 construction schemas that exhibit hierarchical inheritance relationships within themselves. This chapter also presents the linguistic evidences for the conceptualizations underlying the proposed schemas.

Chapter 5 discusses in detail about the proposed annotation framework, experiments, results and error analysis.

Chapter 6 describes the approach that we adopted to build a full parser pipeline. The conclusions and future work that can be extended from here are discussed finally in chapter 7.

Chapter 2

Limitations of the current approaches in parsing Dravidian syntax

Syntactic parsing is the task of automatically finding out the grammatical structure of sentences in a given grammar formalism. Phrase structure grammar and dependency grammar are two such formalisms. Parsing of natural language texts has been explored in various languages for more than two decades now. A syntactic parser implementation can be grammar-driven, data-driven or hybrid. Due to the availability of annotated corpora in many languages, data driven approaches have been successfully employed to develop parsers in multiple languages across the world.

Indian languages are morphologically rich, free-word order languages and it is well accepted that dependency framework suits better for the analysis of the various grammatical structures of such languages[45, 39, 9]. The grammar formalism known as 'Computational Paninian Grammar' (CPG) has been successfully employed to build treebanks in Indian languages[11] and is extensively used for several free word order Indian languages. Prashanth Mannem proposed a bidirectional dependency parser for Hindi, Telugu and Bangla language[37]. Joakim Nivre presented the work of optimizing Malt Parser for three Indian languages namely Hindi, Telugu and Bangla in NLP Tool Contest at ICON 2009[42]. Bharat Ram Ambati et al. explored MALT and MST parsers on three Indian languages Hindi, Telugu and Bangla [1]. Straka et al report an LAS score of 69.7% and UAS score of 78.3% on the Universal Dependencies Treebank[49] for Tamil parsing using 600 sentences. A hybrid approach in 2012 in which the output of both the MALT and MST are combined in an intuitive way showed that this can perform better than both the parsers[32]. In 2011, a constraint based Hybrid dependency parser for Telugu was reported[17], with an LAS score of 68.06% from a training size of 1119 sentences. In Hindi-Urdu treebank project, Bhat et al report a gold LAS score of 92.24% and auto LAS score of 89.19% for Hindi[14]. In 2016, Bhat et al report an improvement of dependency parsing of Hindi and Urdu by modelling syntactically relevant phenomena[13]. For other morphologically rich languages across the world, average LAS scores vary from 91.83% (German language with the largest training set) to around 83% (for Hebrew and Swedish with smallest training data) amongst the languages in SPMRL shared task 2013[44].

The table 2.1 below shows some of the state-of-the-art parsing results reported for a few Indian languages in data-driven approaches. Hindi, Telugu and Bangla accuracies reported are experimented with Computational Paninian Grammar Framework[11]. The Tamil accuracy reported is based on the

Table 2.1 Label accuracy scores(LAS) reported for various Indian languages

Language	LAS	Annotation Scheme	Training size
Hindi	92.24%	CPG	16,629
Telugu	68.06%	CPG	1119
Bangla	79.81%	CPG	1000
Tamil	69.7%	Prague UDT	600

Universal Treebank results reported by Straka et al[49]. Morphologically rich Indian languages (MRLs) express multiple levels of information already at the word level. The lexical information of each word in an MRL may be augmented with other higher level information such as grammatical category of the word, its relation with other words, clitics, particles, inflectional affixes and so on. In English many of these notions are embedded implicitly by word order and adjacency. For instance, a direct object is generally the first NP following the verb in English and it is not explicitly marked in any other way. Expressing these functional ideas inside the morphology of a word allows for larger degree of word order variation since grammatical information need not be associated with syntactic positions.

Although all Indian languages in general are said to be morphologically rich and therefore have relatively free word order, there exist considerable differences among them with respect to their finer morphological properties[30]. While languages such as Hindi has postpositions as case markers, a language like Telugu encodes the case information with suffixes inside the word. Divergences of this kind are easy to handle because the noun and postposition for instance can be locally grouped together and after the local word grouping[11], the representation is similar in both Telugu and Hindi. Hence, the received wisdom is that by employing a good POS tagger, morph analyser and chunker, if a good shallow parsing can be achieved, then already we have a skeleton of the syntactic structure of a language in a common representation format. This shallow parsed representation is extremely useful: (a) In Machine Translation, the source shallow parsed representation can be directly transferred to shallow parsed output of another language and one can proceed with translation from there (b) Learning the syntactico-semantic karaka relations for developing a full parser system is just one step away from shallow parsed representation.

But we notice that despite such a linguistically grounded approach, we see from table 2.1 that even with comparable data size for training and with same grammar formalism used for annotation, the LAS scores are very different for Telugu and Bangla. There are a few reasons for this. Primarily, the Bangla parser was trained on a coarser tagset than Telugu which naturally leads to better learning. Secondly, Telugu is morphologically richer than Bangla with agglutinative properties that makes it difficult to parse. Due to their agglutinative nature, languages like Telugu tend to combine multiple words into one single token, the phenomenon which is called as external sandhi. Kolachina et al[30] point out that external sandhis pose a problem in Telugu parsing. The paper argues that even though the treebank data of Telugu, Hindi and Bangla released in 2009[23] contained relatively comparable data size of 1700, 1800 and 1280 sentences respectively, Hindi has 9.18 chunks per sentence; Telugu shows only 3.78 chunks per sentence. The average number of words in a Hindi sentence is 19.01, 10.52 in Bangla but

just 5.43 words per sentence in Telugu due to the external sandhis. After sandhi splitting and chunking of Telugu data, they report an increase in parser accuracy of Telugu (LAS score of 68.31% on set 3). While the average number of chunks is comparable to Hindi or Bangla, the accuracy is not still comparable to Hindi or Bangla.

In this work we hypothesize that it is not just the morphological richness or the computational complexities that are associated with sandhi splitting, POS tagger and morph analyser which limit the performance of parsing a Dravidian language such as Telugu. It is not just how complex the morphosyntax is but rather what all is encoded by the language through its peculiar morpho-syntax that needs to be understood better.

We are pointing out in this work that there are fundamental facts of theoretical importance that are not considered in the current approaches of Dravidian language syntactic parsers[23, 37, 30, 43]. In the following sections 2.1,2.2,2.3,2.4,2.5 we list down some of the theoretically important syntactic peculiarities observed in Dravidian languages. Subsequently in section 2.6 we show how these peculiarities are not handled in our existing dependency annotation schemes and how this can pose learning problems for a full parser.

2.1 Tense and finiteness

All sentences in Dravidian languages have a series of non-finite verbs followed by only one finite verb at the end no matter how complex the sentence is. In other words, in languages such as Tamil, Telugu you cannot have multiple finite clauses connected by subordinating or coordinating conjunctions unlike Hindi, English or Bangla. The only exception is complement clause constructions where one event is quoted within another event. For example in Hindi, it is perfectly fine to say *maiM dillI gayA aur apnE dOstOm sE milA* 'I went to Delhi and met my friends' with two finite clausal heads *gayA* 'went' and *milA* 'met' connected by a coordinating conjunction *aur* 'and'. But such a construction is infelicitous in Dravidian syntax as shown by Telugu and Tamil examples below.

* *nEnu dillIki veLLAnu* **mariyu** *nA mitrulani kalisAnu* (Telugu)
* *nAn tillikku senREn* **maRRum** *en naNparkaLai santittEn* (Tamil)

A full gloss is not given here for simplicity. We will discuss the constructions in detail in later chapters. For now we only point out that simple coordination of finite clauses makes the sentence infelicitous and awkward. The finite verbs are underlined above and the coordinating conjunctions are shown in boldface. A natural rendering would be to either write them as two different sentences or make the first verb as a non-finite adverbial participle as follows: *nEnu dilliki veLLi nA mitrulanu kalisAnu*. The first verb becomes the non-finite form *veLLi* and only the second verb shows finite inflection as *kalisAnu*.

In these languages, it should be noticed that the non-finite inflections also show the so-called tense markers[5]. For example Tamil verb *vA* 'come' inflects as conditional participial *vantAl* which already shows the past marker *nt* within. *paRa* 'say' in Malayalam inflects as conditional *paRanjAl*, conjunctive *paRanj* with past marker *nj* within. It must be clear by now that tense and finiteness of these languages are peculiar and understanding them is a prerequisite before parsing / generating valid sentences.

2.2 The question of grammatical categories

In addition to tense and finiteness, the following questions are also relevant to Dravidian syntax. Do adjectives and adverbs constitute as basic, distinct grammatical categories? The apparent lack of adjectives has been studied and questioned by Mythili Menon[40], Amritavalli and Jayaseelan[6]. Steever talks about how Dravidian languages by and large lack those parts of speech which are so profitably exploited elsewhere in the study of syntax: conjunctions, complementizers, and adverbs[46]. What kind of grammatical category do pronominalized relative clauses (in the sense as used by Bhadriraju Krishnamurthy[31]) constitute in syntactic analysis?

Simply put, most of the function words that are found in other languages are expressed as verbal inflections in Dravidian morphology. For instance, instead of subordinate conjunctions that link two clauses we find non-finite adverbial inflections in all the clauses except the main clause. Another example is the case of postpositions that are expressed through conjunctive participial inflection of verbs. Not only that; these postpositions are not simply syntactically frozen but may further inflect as relative participles or as pronominalized relative clauses. Take the Telugu postposition *gurinci* 'about'. This may further inflect as *gurincina*, *gurincinadi*, *gurincinavi* and so on just like it were an actual verb while still being treated as a postposition. Same is true in Tamil as well, as shown below.

paRRi about (postposition derived from conjunctive form of verb *paRRu* - 'to hold / stick')
paRRiya (that) which is about (Relative participial form of the verb *paRRu*)
paRRiyatu that thing which is about (Relative participial pronominalization of *paRRu* with non-human singular suffix)
paRRiyavargaL those people who are about (Relative participial pronominalization of *paRRu* with human plural suffix)

These verbal inflections holds good for adjectives, adverbs, quotative markers etc. Why do verbal inflections occur in these function words? How to understand and process them in a principled way is a question that needs to be addressed before we do shallow parsing.

2.3 Similar patterns in different syntactic environments

The third crucial characteristic of Dravidian syntax is the occurrence of same morphological forms in formally different syntactic environments. Consider a sentence in English 'I went to canteen and ate idlis'. As has already been discussed, one cannot write two finite clauses and coordinate them in Dravidian languages. Therefore the equivalent constructions in Telugu and Tamil, for instance, are shown below:

(1) *nEnu* *kyAnTeen-ki* **veLL − i** *iDlI-lu* *tin-n-Anu*
I canteen-DAT go-CONJ idli-PL eat-PST-1.SG
'I went to canteen and ate idlis'

(2) *nAn* *kAnTeen-ukku* **pO − y** *iTlI* *sAppiT-T-En*
I canteen-DAT go-CONJ idli eat-PST-1.SG
'I went to canteen and ate idlis'

It can be seen that instead of a coordinating conjunction, the first verb takes a conjunctive participial inflection i.e. *pO* 'go' inflects as *pO-y* to build a grammatically correct sentence. This conjunctive participial inflection is seen in other Indian languages as well but in Dravidian languages such non-finite inflections seem to be the only grammatical choice available. It is an observed fact that a typical Dravidian sentence exhibits a series of non-finite verbs followed by one finite verb at the end (being head final languages) [5, 46]. In the absence of a distinct inventory of subordinate conjunctions, these languages heavily rely on their non-finite participial inflections. But, interestingly, the same non-finite inflections which occur between multiple clauses are also observed in monoclausal syntactic environments such as grammatical aspects, modalities, serial verb constructions etc. Look at the below examples in Telugu and Tamil.

(3) *rAmuDu* **vacc − i** *un-T-ADu*
Ram come-CONJ AUX-FUT-3.M.SG
'Ram would have come'

(4) *rAman* **va − ntu** *iru-pp-An*
Ram come-CONJ AUX-FUT-3.M.SG
'Ram would have come'

Comparing examples 1,2 and 3,4, it can be seen that the same conjunctive participial forms which occurs in event sequences appear in constructions expressing grammatical aspects. In formal syntactic analysis this participial inflection in grammatical aspect is considered as a readily available grammatical configuration[8]. The same non-finite inflections occur in various different function words such as adjective, adverb, postposition etc through grammaticalization. We suggest that there are meaningful discourse conceptualizations underlying these non-finite inflections and it is by understanding the principles behind these non-finite inflections that we can formulate Dravidian language parsing better.

2.4 Ambiguities of morphosyntactic forms

Not only do Dravidian languages rely on non-finite participial inflections to express relations between clauses, but also they are ambiguous on their surface morphological forms. Look at the table 2.2 below. Table 2.2 shows a few morphological forms highlighted in bold that give rise to various interpretations

Table 2.2 Possible interpretations of same morphological form

Example	**Meaning**
vanta paiyan	the boy who came
vanta Ur	the village to which (one) came
vantapozhutu	when one came
kOpam **vanta** mAtiri	as if anger came, just like how anger came etc.
kOpam **varukiRa** mAtiri	such a way that anger comes, as if anger comes, similar to how (one) gets angry etc.
nInkaL **vantatu** tirupti	the fact that you came is satisfactory
vantatAl muTintatu	Because (someone) came, this was possible
nI **vantatu** teriyum	(I) know that you came
avan **vantatai** pArttEn	I saw him coming
vantatu vITTukku	It is to a house that (one) came

in various contexts. A full gloss of every morpheme is not given as it is not indispensable to what we want to point out. The relevant surface form is highlighted in bold. As you can see, the same word *va-nt-a* 'come-PST-RP' with relative participial (RP) inflection is used to mean participant, the manner of action, the time of action, the resultant of discourse attribution etc based on the semantic constraints between this RP form and the noun modified by it. Unlike English, there are no structural markers such as relative pronouns, complementizer or prepositions that relate the verb with the noun; there are no subordinate conjunctions such as *when, as* etc. Likewise the same word *vantatu* is used to mean *the one who came, the place to which one came, the fact that someone came, the act of one's coming* etc. It should be observed that the constructions like *vantatu* are not mere morphological fusion of the RP form *va-nt-a* 'come-PST-RP' and a simple pronoun like *atu* 'it'. i.e. *vantatu* $\neq$ *vanta+atu*. The ungrammaticality of the example 6 below demonstrates the fact.

(5) *rAman va-nt-a-t-ai nAn pAr-tt-En*
Ram come-PST-RP-NH.PRON-ACC I see-PST-1.SG
'I saw Ram coming'

(6) **rAman va-nt-a at-ai nAn pAr-tt-En*
Ram come-PST-RP it-ACC I see-PST-1.SG
'*I saw Ram coming'

I want to point out that the morphological information and syntactic categories seem often at odds with each other and most such forms are ambiguous within their own clause. In fact under different sentential contexts the larger syntactic function of an expression turns out to be very different from what you would expect from its given morphological form even if the whole clausal context is available. The following examples illustrate the point.

(7) *pOna vAram kET-T-a pATal-ai* $\underline{allA-mal}$
last week hear-PST-RP song-ACC negate-CONJ
nERRu kET-T-a pATal-ai tAn virumpu-kiR-En
yesterday hear-PST-RP song-ACC only like-PRES-1P.SG

'I like the song that I heard yesterday <u>and not</u> the one I heard last week'

(8) *nERRu kET-T-a pATal-ai* $\underline{allA-mal}$ *vERu*
Yesterday hear-PST-RP song-ACC negate-CONJ other
et-ai-yum nAn virumpavillai
which-ACC-also I like-NEG

'I do not like any other thing <u>except</u> the song I heard yesterday'

In the above examples 7 and 8, the same non-finite form *allAmal* performs two different functions. In the former it is understood as connecting two clauses but in the latter it functions like a postposition. This is because, we say, these verbal inflections have discourse meanings and they are understood from how they interact with other clauses. From a processing point of view, it appears that only in the context of other clauses and their semantic constraints that the formal syntactic function itself becomes determinable.

2.5 Semantics of particles

The particles such as *um, O, A* etc. have their own syntactic and semantic properties which need to be understood and incorporated in parser learning. As an example, the particle *um* is treated as a coordinating conjunction in our current approaches to parsing. In fact the particle *um* does not function like a coordinating conjunction but like a list operator that takes as its input a list of grammatical units and creates dependency with a verb as the head of this list. It follows therefore that *um* cannot be added to, say adjectives, relative participial forms, genitive form of a noun, oblique form of a noun etc. Look at the ungrammaticality of the below constructions when we try to add these particles to units which modify a noun.

- ennuTaiya(my) pEnA(pen) - My pen
- *ennuTaiyavum(my-operator) pEnA(pen) - Pen of mine as well.
- vanna(come-pst-RP) kuTTi(child) - The child which came
- *vannavum(come-pst-RP-operator) kuTTi(child) - The child which also came
- andamaina(beauty-become-RP) mukham(face) - The beautiful face
- *andamainE(beauty-become-RP-operator) mukham - The face which is indeed beautiful

Understanding this phenomenon and treating them as operators instead of morphological equivalents of conjunctions, disjunctions etc. can result in significant improvement in handling peculiar constructions in Tamil.

2.6 Computational issues

Due to the theoretical issues mentioned in sections 2.1,2.2,2.3,2.4,2.5 many learning issues occur in current approaches of Dravidian language syntactic parsers. We will just point out to some such learning issues. Look at the example 9 below.

(9) *nINT-a-t-um* *kaLaippu* *mikk-a-t-um-An-a*
elongate-RP-PronSuffix-also tiredness exceed-RP-PronSuffix-also-became-RP
payaNam *toTar-nt-atu*
journey continue-pst-agr

'The long and arduous journey continued'

As it can be seen there are no pure adjectives in Tamil and therefore an expression like *long and arduous* is rendered by a complex expression *nINTatum kaLaippu mikkatumAna* whose literal morphological glosses are given above. One of the ways in which the parts of speech and chunking of the above sentence can be done in the current approaches is as follows:

1. ((nINTatum NN)) - Noun chunk
2. ((kaLaippu NN)) - Noun chunk
3. ((mikkatumAna JJ) (payaNam NN)) - Noun chunk
4. ((toTarntatu VB)) - Verb chunk

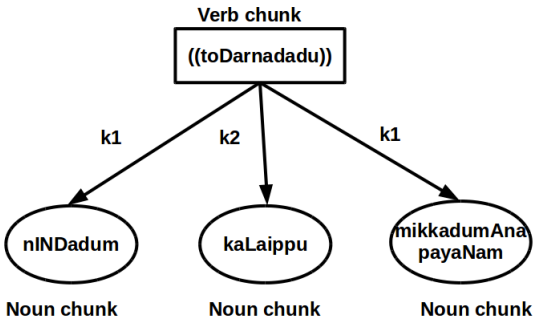


Figure 2.1 Wrong dependency that is likely to be learnt

Note that the expression 'nINTadum' can be treated in two ways: (i) as a noun (NN) and in turn as an NP chunk because morphologically it exhibits the property of a noun with pronominal suffix or

(ii) as a main verb (VM) and in turn as a VGNF chunk because it shows relative participial verbal inflection. ‘mikkatumAna’ is tagged as an adjective (JJ). However, since chunk is structurally defined in CPG as minimal non-recursive unit of analysis[12] the second expression ‘mikkatumAna’ and the noun ‘payaNam’ together are by definition grouped together as one Noun chunk. Now once it is chunked like this, there is no way to learn that there are two expressions that are modifying the word ‘payaNam’. If the first ‘nINTatum’ is treated as NP chunk the most likely dependency tree that the parser will end up learning is as shown in figure 2.1.

Even if ‘nINTatum’ is treated as a VGNF chunk the system fails to learn that it has dependency with ‘payaNam’. It is more likely that such a VGNF chunk becomes a dependent with the last verb chunk *toTarntatu*. This is a type of error that cannot be handled by just increasing the training data size because the problem here has a theoretical basis where there is a mismatch between what the morphology of the language says and what formal syntactic category the word stands for.

Yet again, look at the following scenario where the particle like *um* with the ambiguous non-finite verbal form *Aka* poses a learning issue.

(10)

avan nIr iRai-pp-a-t-um avaL tI(y)-ai mUTTu-v-a-t-um-Aka
He water fetch-FUT-RP-PRON-OPER she fir-ACCe light-FUT-RP-PRON-OPER-RB
vElaikaL-ai pakir-nt-u koNT-anar
work-ACC share-PST-CONJ REFL.AUX-AGR.

‘They shared the tasks between themselves such as he fetching the water and she lighting up the fire (for cooking)’

Let us look at the learning issues that invariably occur in these kind of constructions. Example 10 is usually chunked as follows:

1. ((avan PRP)) - NP chunk
2. ((nIr NN)) - NP chunk
3. ((iRaippatum VM)) - VGNF chunk
4. ((avaL PRP)) - NP chunk
5. ((tIyai NN)) - NP chunk
6. ((mUTTuvatumAka RB)) - VGNF chunk

The morphological unit *Aka* within the second underlined verb *mUTTuvatumAka* cannot be treated formally as an adverbializer to tag the whole expression as an adverb (RB); Nor can you treat the unit *Aka* as a verb because it does not indicate any event but simply behaves as a function word. The peculiarity is non-finite verbal inflections are grammaticalized as function words but are morphologically still verbs and hence *um* can take a set of words and list them as dependants of this function word. Take another example.

(11) arasAnkam son-n-a-t-ai(y)-um avarkaL sey-t-a-t-ai(y)-um
Government tell-PST-RP-PRON-ACC-OPER they do-PST-RP-PRON-ACC-OPER
paRRiy-a vivAtam
regarding-RP debate

'Debate about what the government told and what they did...'

Again, in example 11, if the underlined non-finite verbs with pronoun suffixes are treated as VGNF chunks, then we end up treating the postposition *paRRiya* as a separate VGNF chunk and it loses the intended function of postposition even during shallow parsing. If the underlined units are treated as NP chunks due to pronominal suffixes and then the expression *paRRiya* is tagged as postposition (PSP), it becomes a part of the second NP chunk but the parser will eventually fail to learn that *paRRiya* is a noun modifier for the word *vivAtam* 'debate'.

The point is that since *um* is a list operator and not a mere conjunction, it attaches to various word classes and creates dependencies with morphological verbs as the list head. However the morphological verbiness does not translate as syntactic verb since most function words are verbal inflections in Tamil and most of these surface forms seem ambiguous. Our current chunked representation assumes that once the formal syntactic category of a word is identified and its morphological information is captured and chunking is done, the shallow parsed syntactic skeleton is available for richer processing. Dravidian languages pose hindrances to this approach in that the formal syntactic function of many expressions can be determined only when taken together with other clauses and semantic constraints.

Our observation is that most of the morphological inflections and formal syntactic functions that arise therein are in fact based on a finite number of meaningful discourse conceptualizations. It is these discourse concepts that are encoded by the morpho-syntactic constructions in these languages. Our larger thesis is that by systematically formulating the relation between discourse conceptualizations and surface morphological forms, we will be in a better position to process the syntactic structure of these languages. We suggest that instead of a formal generative approach to syntax, functional approaches such as Construction grammar provide theoretical possibilities to understand and explain Dravidian syntax better. We will discuss this approach in chapters 3 and 4.

Chapter 3

Construction grammar and basic schemas

In contrast to the paradigms of formal linguistics, which sees meaning as less relevant or sometimes autonomous [15], functional approaches formulate grammar from a meaning-oriented perspective. Functional theories analyze the grammatical structure of a language as do formal theories; but it also analyzes the entire communicative situation: the purpose of speech event, participants, discourse context etc[41]. Thus there are various schools of functional approaches to grammar such as Prague Functionalism[36, 24], Functional Discourse Grammar[21], Systemic Functional Grammar[19], Role and Reference Grammar[51], Construction Grammar[20][16], Cognitive Grammar [34] and so on based on what type of functional analysis they engage in. Amongst the functional approaches, cognitive linguistic enterprise marks a school of thought and practice where language is treated as an integrated part of human cognition which operates in interaction with and on the basis of same principles as other cognitive faculties[18]. There are various strands or orientations to this paradigm that gives rise to different theories such as gestalt-psychology based strand, phenomenology based strand, a cognitive discourse based strand, cognitive sociolinguistic strand and psycholinguistic strand[18]. We take inspiration from construction grammar and cognitive grammar approaches to language which subscribe to the idea that knowledge of a language is based on a 'collection of form and function pairings'.

Under such a view, language is understood as a symbolic system all the way from lexicon to grammar. It follows that the distinctions between lexicon, morphology, syntax are just the differences in the schematic complexity of the symbolic expressions, but symbols nonetheless. The basic units of analysis are *symbolic expressions* or commonly *constructions*. For example just like you learn the lexical item *dog*, in English, to mean the concept of a dog, so do you learn a construction like *The more the X the more the Y* to mean something like 'Given two predicates X and Y, Y is said to vary on the scale of comparison in proportion to X'. Just like the lexical item *dog* is arbitrarily a symbolic association with the concept it represents, so is the construction an arbitrary symbolic association with the concept mentioned above. The same concept of proportional increase is expressed in Tamil by the construction 'How much to how much X, that much to that much Y' which shows that even the schematic expressions are arbitrary in various languages by virtue of their being symbolic. Refer to [35] for a detailed understanding of how grammar can be treated as a symbolic system. Based on this understanding, we

proceed to understand the basic grammatical categories that make up a language. For our purposes, we are not taking into consideration the subtle theoretical differences between Construction Grammar and Cognitive Grammar, but focus on their common theoretical ground relevant for our task.

3.1 Grammatical categories as understood in Cognitive Grammar

In elementary schools, it is taught that noun is a name of a person, place or a thing. Later as a linguistics student, one learns that a grammatical class like *noun* can only be defined in terms of its grammatical behaviour and that the conceptual definitions of grammatical classes appear to be impossible[33, 25]. Most of the arguments against a conceptual definition of grammatical categories is based on the objectivist semantics, where an expression's meaning is identified with the objective features of the situation described. Hence, conventionally it is argued that *redness* being an attribute, *earthquake* being an action, *dog* being an animal, *silence* being an experience and so on, there is no systematic way to define *noun* semantically. But such an objectivist account of meaning fails to take into consideration the capacity of human beings to cognitively view the same situation in alternate ways. Look at the sentences below which are objectively same from an information point of view but cognitively produces different viewpoints of the same information.

- The glass contains water
- The water is inside the glass

The difference between the sentences does not lie in their information content but in its construal content. The cognitive ability of the human being to conceptualize a given situation in many different ways is called by Langacker as 'construal'. Expressions like *explode* and *explosion* refer to an event objectively but a person construes the latter expression as an abstract thing through conceptual *reification*.

In this approach, grammatical patterns are analysed in terms of *Construction Schema*. A *construction* is defined as either an expression (of any size) or a schema abstracted from expressions to capture their commonality to any level of specificity. Thus construction schemas are set of meaningful templates that define some semantic conceptualization and any expression which inherits the same conceptualization with more specificity is said to instantiate the construction schema. The specific construction that is generated out of the schema is called its *instance*. A construction schema itself can be inherited from other schemas and define additional concepts as a part of its definition. Hence one schema can inherit properties from multiple other schemas and an instance of such a schema exhibits all the properties inherited from these multiple schemas.

Basically, the theoretical idea is that an expression's meaning not only depends on the information content but also on what kind of viewpoints or perspectives that a speaker takes while 'languaging'. The conceptualizations dynamically made by the speaker are called 'construals' and the syntactic patterns that act as symbolic mappings to these construals are called construction schemas. It has a phonological pole and a semantic pole by definition of being a symbol. The construal behind a construction schema

is characterized by the level of *specificity*, the *perspective* or *viewpoint* and the *degree of prominence* conferred on the elements within a construal. I am not elaborating every dimension listed above but take the most relevant concepts based on which further concepts relevant for Dravidian syntax can be build upon.

Any linguistic expression triggers some kind of conception as the basis for its meaning. Out of this conceptual base, one sub-portion stands out as a focus of attention and it is this one that the expression refers to. For instance, if you utter the word *arc*, it invokes a conceptual base of a *circle* and a section/portion of it is focussed as the referent of the expression. The expression *roof* may invoke an idea of a 'house' and a sub-part of it is referred by this expression. This subsection that a linguistic expression refers to in the overall construal of an idea is called as its *profile* in CG terms. Any expression in its most abstract schematic sense can profile either a *thing* or a *relationship*. CG uses the word 'entity' as a generic term to refer to any idea that is available for description by grammar.

The construal of a *thing* invokes **Noun schema**. A *thing* is any "region in some domain", a product of grouping and reification. Words like *boy, recipe, committee* and so on are not necessarily unitary objects by themselves but their constituent features are cognitively grouped and reified as a single unit which makes them a noun. For example when you look at the sky and observe a set of stars and call it a 'Big bear', obviously it is not meant that the stars are objectively connected as one unitary unit spatially but rather the constituent stars are grouped and reified as one entity at a higher level of conceptual organization. Thus by the grouping and reification of constituent entities we invoke Noun schema.

The construal of a *relation* invokes **Verb schema**. A *relation* is the notion of viewing an entity not in isolation but as a part in the overall scheme of configuration[35]. Verb schema presupposes two fundamental cognitive abilities: the capacity for apprehending relationships and for tracking relationships through times. Here again the relationship involves grouping of entities in some configuration, but instead of *reification* of the grouped entities if we focus on the *interconnections* themselves, we invoke the Verb schema. Take for instance three words: *into, enter* and *entry*. To conceive the meaning of these words, two entities form the conceptual base and the following three profiles are possible as shown in figure 3.1. From these image schema representations, one can see that when the focus is on the config-

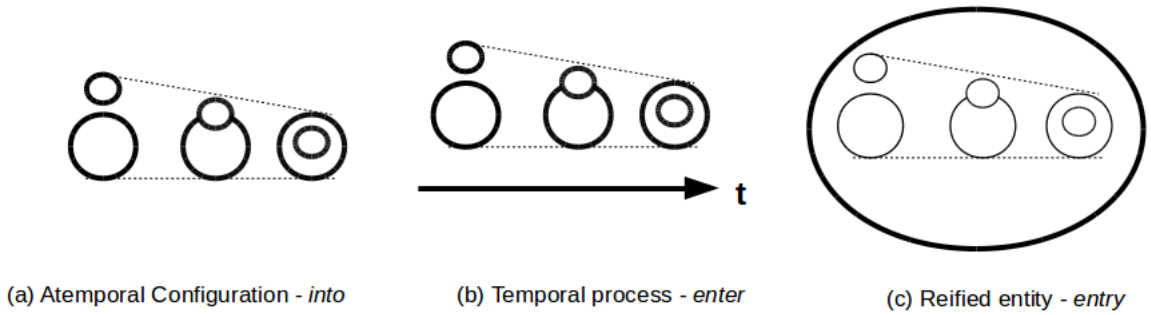


Figure 3.1 Construals underlying the same configuration

uration itself atemporally it corresponds to the preposition interpretation of *into*, when the focus is on

the same configuration coming into being temporally through mental scanning, it gives rise to the verbal interpretation of *enter* and when the focus is on the reification of the entire group that makes up this configuration it gives rise to the noun interpretation of *entry*.

There are other concepts like *trajectory, landmark, vantage point* etc. relevant for CG, but for our purposes this introduction is sufficient. We will point out some construction patterns observed in Dravidian languages and proceed to formulate our own insights about what are the construction schemas that make up Dravidian grammar. Our analyses primarily rely on data from Telugu and Tamil but we compare constructions from Malayalam and Kannada as and when possible.

3.2 Construction Grammar basics for Dravidian languages

Two properties of Dravidian languages offer us clues to two dimensions of understanding any construal. The two properties are:

- Similar morphological inflections occur in formally different syntactic environments
- Every morphosyntactic form creates a sense of completeness or incompleteness in the larger discourse context

The first point gives us a clue that there must be something common in all these formally distinct syntactic environments that these languages have chosen to encode similar inflections in all these cases. In other words, regardless of the conceptual *base* from which an expression arises in all these syntactic environments, there must be a common *profile* that all these seemingly different syntactic units invoke that needs to be discovered. This dimension of analysis of a construal via its *profile*, leads us through more and more abstract schemas from which the current *expression* has inherited all its properties and ultimately is reducible to either a *thing* or a *relation*. This dimension is what we choose to call as *compostion* – what all schemas through their hierearchical inheritance relationships make up an expression.

The second point gives us clue towards the other dimension of analysis of a construal. While conceiving an expression, we not only construe about its *profile* (and in turn its whole composition) but also about the expectation that we have out of this expression in terms of its potential to build the discourse. This second dimension of analysis is what we call as *interaction*. An expression not only has the property of being *composed* of some schemas but also the potential of *interaction* with other such schemas. This expectation of interaction should be satisfied by some specific expression whose schematic composition complements these expectations.

As an example, take the expression *nIN-T-a* 'elongate-PST-RP' in Tamil which has the following properties: It can be interpreted as an adjective 'long' as in 'nINTa payaNam' to mean 'long journey'. It can also be interpreted as the event of 'elongation/ stretching' as in the expression 'iTuppu varai nINTa avaLatu kUntal' to mean 'her hair which stretched upto her waist'. How such interpretations can be understood schematically is one dimension of analysis; there is the other dimension of analysis. When somebody utters an expression 'nINTa', an expectation of a *thing* is created in discourse, not a

relation. Hence *nINTa pAtai* 'long path', *nINTa kUntal* 'long hair', *nINTatu* 'the thing which is long/ that thing which stretched ' etc are valid expressions but not * *nINTa senRAn* '* He went *nINTa*' or **nINTa muTintatu* '* It was possible *nINTa*' etc. No sense could be made out of such expressions since the discourse expectation underlying the RP inflection is a *thing* while these expressions do not provide one.

The basic idea is that every construction is construed not in isolation but within a larger discourse world. Thus constantly, while uttering an expression we construe that it is *made up of some schemas* and it *expects some schemas*. This expectation and satisfaction continues till the speaker reaches an expression which she considers as revealing the nuclear discourse idea (for which the discourse world is created mentally in the first place) and once this nuclear discourse unit is expressed, the sentence/ utterance is said to be *complete*. At any given point of creating/ analysing the discourse, the expression which satisfies a discourse expectation becomes the head and the expression which creates a discourse expectation becomes the dependent. Since all other expressions except the nuclear discourse idea constantly build more expectations in the discourse, every expression except the nuclear idea is a head and dependent of some other expression.

We hypothesize that the morphological inflections observed in Dravidian languages are directly encoding these dimensions of information on every expression. On the outset, we propose that *composition* and *interaction* are two dimensions for understanding any syntactic unit functionally as a construction schema. At the most abstract schematic level, every construal is either a thing or relationship. A thing can be construed to expect another thing or a relationship. Similarly a relationship can be construed to expect another thing or a relationship. Hence there are four types of interactions possible which is shown in figure 3.2. The direction of the arrow points to the dependent unit in the interaction. To make

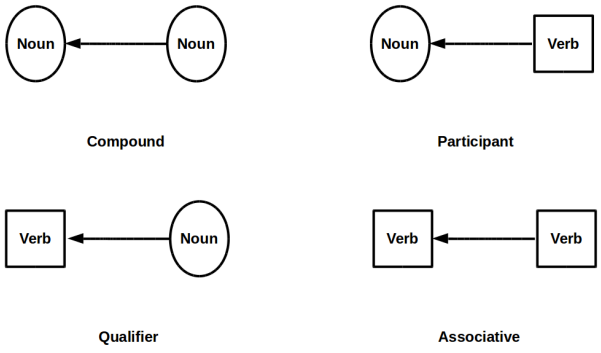


Figure 3.2 Basic Types of Interactions in Discourse World

a fully functional characterization of syntax, the following construction schemas are defined based on the conceptions they invoke in the discourse world.

1. Any morphosyntactic unit that is conceived as outlining a *thing* irrespective of its internal complexity is a 'noun' schema. Anything which is conceived as outlining a *relation* is a 'verb' schema.

2. A relation can be conceived as outlined through time or immediately perceived as a configuration. We refer to the former interpretation as a ‘process’ and the latter as a ‘status’.

3. If a relation attributes its relational properties on a thing, it is called as ‘qualifier’ schema

4. If a relation attributes its relational properties to another relation , it is called as ‘associative’ schema

5. If a thing is conceived to be an integral part of another thing, it is said to be in ‘combinative’ schema

6. If a thing is conceived to be an integral part of a relation, it said to be in ‘participant’ schema

The above conceptions form the basic schemas whose hierarchical relationships are portrayed below in figure 3.3.

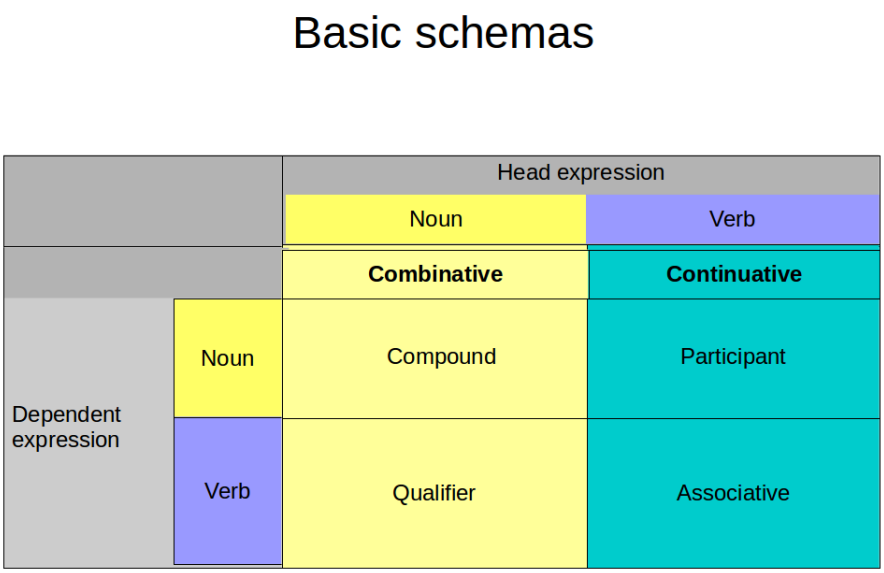


Figure 3.3 List of basic schemas

We will discuss in detail all these schemas with relevant examples along the two dimensions: *composition* and *interaction*.

3.3 Schemas along the dimension of composition

3.3.1 Noun Schema

A noun is a construction schema which outlines a *thing* as an instance. A *thing* is any “region in some domain”[34]. For example *paiyan* ‘boy’, *nallavan* ‘the one who is good’, *vantavan* ‘the one who

came' are expressions of various schematic complexities but in each case a *thing* is conceived by the speaker eventually (regardless of what other internal concepts these expressions represent). All these are said to be inherit a *noun* schema eventually.

3.3.2 Verb Schema

A Verb is a construction schema which outlines a *relation* as an instance. A *relation* is the notion of viewing an entity not in isolation but as a part in the overall scheme of configuration.(Inspired from Cognitive Grammar[35]). When such a relation is exhibited as if it is outlined through time, it is called as *process*. When such a relation is exhibited as if it is outlined instantaneously as a configuration, then it is called as *status*. We use the word *status* to mean all atemporal relations that are linguistically possible. Both process and status inherit the verb schema because of the common abstract property of outlining a relationship. The above Cognitive Grammar interpretation cannot be more explicit than in the case of Dravidian languages which actually use the same verb morphology for atemporal relations as well.

For example, a surface form like *nINTa* 'elongate-PST-RP' has a property of bringing out a relation and hence is a verb. This surface form can have two interpretations: *the one which became long* or *the one which is long*. Now these two are conceptually different. The former is a 'process', which describes a relation of length variation profiled through some time. In the latter interpretation, it is a 'status' which describes the relation of length variation as a configuration profiled instantaneously. Let us discuss about these two viewpoints.

3.3.3 Process and Status Schema

Look at the examples 12 and 13 below.

(12) en kaN munnE **nIN-T-a** anta pAtaiy-ai
my eye before elongate-PST-RP that path-ACC
nAn kaN-Tu rasi-t-tEn
I see-CONJ enjoy-PST-1.SG

'I enjoyed looking at the path **which stretched** before my eyes'

In the above example, 'nINTa' profiles a process of elongation/ stretching which can be thought of as a sequential scanning through some time.

(13) atu oru **nIN-T-a** pAtai
Dem.Pron one elongate-PST-RP way/path
'That is a long path' (Status)

'That is a path which stretched out' (Process)

nINTa in this case profiles a configuration of *already having become long* at the instance of description and not as a process which takes some time for sequential scanning. Therefore it denotes a *status*. Thus

we see that the relative participial (RP) inflection of a verb such as *nIL* 'elongate' has two interpretations depending on how the speaker perceives the relation.

When a relation is sequentially scanned through mentally as if it takes time for the relation to unfold, that can be considered as a process i.e. an event. The image schema for such a viewpoint is shown in figure 3.4. When the same relation is mentally perceived through summary scanning as if it is just

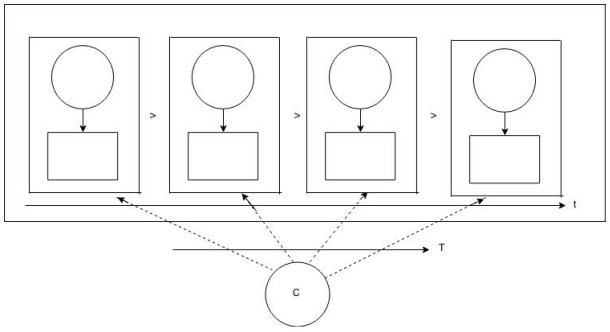


Figure 3.4 Process - Sequential scanning of how the relation unfolds

a totally available configuration rather than a temporal unfolding, such as viewpoint is considered as status i.e. an arrangement/ a configuration. The image schema for this perspective is shown in figure 3.5.

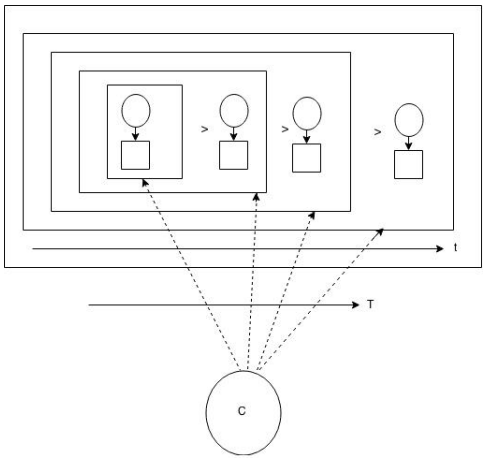


Figure 3.5 Status - Summary scanning of how the relation is configured

It is this idea that is exploited by Dravidian languages to create its adjectives. Morphologically, adjectives in Dravidian languages are usually either (a) a direct RP inflection if derived from verb stem or (b) an RP inflection of the basic verb 'become' added to some noun stem. The adjectives 'fast, beautiful, long, deep' are rendered respectively in Telugu as *vEgam-ai-n-a* 'speed-become-PST-RP', *andam-ai-n-a* 'beauty-become-PST-RP', *poDuv-ai-n-a* 'length-become-PST-RP', *lOt-ai-n-a* 'depth-become-PST-RP' expressed through relative participial (RP) form of basic verb *agu* 'become'. Many other adjectives are directly RPs of verbs such as these Tamil examples: *uyar-nt-a* 'raise-PST-RP' for 'great', *akan-R-a*

'expand-PST-RP' to mean 'wide', *siRa-nt-a* 'excel-PST-RP' to mean 'excellent', *miku-nt-a* 'exceed-PST-RP' to mean 'much' etc.

These morphological inflections functionally trigger the same *process/status* perspective shifts as discussed in example 12 and 13. Under a *status* viewpoint, the surface form is interpreted as an adjective while under a process interpretation the same surface form, for instance, 'akanRa' means 'that which widened/expanded', an actual event. Thus, the speaker shifts his perspective of whether to interpret the given verb as process or status depending on the larger discourse context and semantic constraints of other construction units in the sentence.

Such interpretations are not restricted to abstract attribute nouns like 'calmness','sweetness' etc. The same holds good with other nouns as well. By adding *Ana* to non-attribute non-abstract entity like 'tAy(mother)' the surface form that emerges is 'tAyAna'. Here again we get two interpretations of process and status. As a process we mean 'she who became a mother' but under status interpretation the same word means 'She who is a mother'. Look at the below Malayalam example.

(14) **ammay-A-y-a** avaL van-n-u
mother-become-PST-RP *she* *come-PST-FIN*
'1. She, **who became a mother**, came'(Process)
'2. She, **the mother**, came'(Status)

As such, the surface form is ambiguous because the morphemes that are combining are the same and the inflections are also same. But their schematic interpretations are different based on the constraints imposed by the larger discourse context.

3.4 Function words are status verbs

We had earlier mentioned that Dravidian languages lack a whole lot of distinct function words like subordinate conjunctions, complementizers, adverbs and rather rely totally on the non-finite verbal inflections. Whatever extant syntactic units that can be called as distinct function words are indeed contemporary verbs in non-finite forms. Postpositions, complementizers, adjectives, adverbs, quantifiers etc. are all actual verbs showing non-finite verbal inflections. This only makes sense now knowing that function words are symbolic units that establish atemporal relations / configurations between various entities in a discourse. The relation is available as an immediate configuration and is not conceived as unfolding through time. To that extent all function words are indeed status verbs. It is this status interpretation of verb that is triggered by the Dravidian verbal inflections.

For instance the words for prepositions *about, around, near, beyond* are verbs morphologically, such as *kuRicc(in Malayalam), cuTTU(in Telugu), kiTTa(in Tamil), mIri(in Kannada)* respectively. All these are adverbial participial inflections of actual verbs *kuRi, cuTTu, kiTTu, mIru* respectively. Another example: *tANTi* 'beyond' is derived directly from a verbal stem *tANTu* 'to cross' and takes a conjunctive participial inflection to form *tANTi* 'having crossed'. As a *process* the word *tANTi* means '[somebody]

having crossed [something]' but as a *status* the same morphological form means 'to be in a configuration of having crossed' i.e. 'beyond'. Look at the below example.

(15) ravi maratt-ai **tANT-i** pO-n-An
Ravi tree-ACC cross-CONJ go-PST-3.M.SG
'1. Ravi **crossed** the tree **and** went' (process interpretation)
'2. Ravi went **beyond** the tree' (status interpretation)

The word 'tANTi' in the above example is ambiguous in Tamil. It means '[Ravi] having crossed [the tree]' in a *process* interpretation and means 'beyond [the tree]' in status interpretation. Only larger discourse context disambiguates what kind of perspective the reader should hold while reading the sentence. Another example is shown below:

(16) maram Ur-ai **tANT-i** iruk-kiR-atu
Tree village-ACC cross-CONJ.P exist-PRES-3.NH.SG
'The tree is **beyond** the village'

In example 16, the word *tANTi* means 'having crossed (be in that configuration)' which is what 'beyond' means. Such a *status* interpretation is readily available seeing that semantically the process interpretation of 'crossing' is not applicable to the tree which does not move. Hence the semantic constraints imposed by the entities or a larger discourse context allows the listener / reader to shift his schematic interpretation as a process or status.

The same is true for most other postpositions such as *nOkki* 'having looked' to mean 'towards', *taLLi* 'having pushed' to mean 'away', *suRRi* 'having surrounded' to mean 'around' etc. Further, it is observed that these postpositions also inflect for relative participial(RP) forms when they act a noun modifiers, and further inflect as relative clause pronominals(RP-PRON) when they act as referring expressions. The table 3.1 below illustrates the point. From the table 3.1, it can be noted that postpositions which are

Table 3.1 Postpositions as status verbs

Morphosyntactic form	Gloss	Literal meaning	Syntactic function
nOkki	look-conj.part	having looked	towards
nOkkiya	look-conj.part-RP	which, having looked	which is towards
nOkkiyatu	look-conj-part-RP-pron	that thing which having looked	that thing which is towards
nokkiyavarkaL	look-conj-part-RP-hum-pl-pron	the people who, having looked	those ppl who are towards
iTaiyil	midportion-locative	in the middle portion	between
iTayilAna	midportion-loc-become-RP	which is in midportion	which is in between
suRRiya	surround-conj.part-RP	which, having surrounded	which is around
suRRiyavai	surround-conj.part-RP	those things, which having surrounded	those things which are around

morphologically verbal stems inflect for RPs and RP-pronouns directly (What these RP inflections and pronominal inflections conceptually stand for will be discussed in section 3.5.5 and 3.5.7). Where these function words are not morphological verb stems, a default verb 'become' is added to the stem and then the this word inflects now for participial and pronominal forms. *nOkki* 'towards' is derived from verb stem *nOkku* 'look' and directly inflects as *nOkkiya, nOkkiyatu, nOkkiyavai* etc. But the word *iTaiyil* 'in

the middle' is from a nominal stem meaning 'middle' and therefore the stem *Aku* 'become' is added to form *iTaiyil-Aku* and then it inflects as *iTaiyil-Ana, iTaiyil-Anatu, iTaiyil-Anavai* etc. just like any other verb. Other function words such as complementizer 'that', quotative marker, conjunctions such as 'yet',

Table 3.2 Other function words for illustration

Morphosyntactic form	Gloss	Literal meaning	Syntactic function
niRaiya(status)	get-fill-concurrent	as [something] gets filled	many/much
niRaiya(process)	get-fill-concurrent	as [something] gets filled	As/while getting filled
taLLi(status)	push-conjunctive	having pushed [something]	away
taLLi(process)	push-conjunctive	having pushed [something]	Having pushed, then..
AnAlum(status)	become-conditional-also	if [something] became, then also	Yet
AnAlum(process)	become-conditional-also	if [something] became, then also	Even if [X] happened
enRa(status)	say-RP	[something] which [one] said	complement clause relativization
enRa(process)	say-RP	[something] which [one] said	that which [someone] told
surunka(status)	shrink-concurrent	as [something] gets shrunk	shortly
surunka(process)	shrink-concurrent	as [something] gets shrunk	as [something] shrinks
amaitiyAka(status)	silence-become-concurrent	as silence becomes	silently
amaitiyAka(process)	silence-become-concurrent	as [someone] becomes silent	as one becomes silent

'but', 'however', 'although', 'eventhough' , quantifiers such as 'many', 'much', 'less', 'more' etc. are also similarly non-finite forms of verbs. Adverbs in Dravidian languages are again created by participial inflections of verbs if the stem is morphologically verb, otherwise the verb *Aku* 'become' is added as a morpheme and then inflected to participial forms. The table 3.2 shows some examples of the same morphological form in process and status interpretations. The table also points out to the recurring pattern in Dravidian languages that the same verb root is exploited to create a process interpretation and status interpretation. We see the word 'silently' derived from the noun stem 'silence+become-concurrent participial form'. The adverb 'shortly' is derived from the verb stem 'shrink' which inflects as concurrent participial form. Just like adjectives, these kind of adverbial derivations are not restricted to just abstract properties such as 'shortness, slowness, fastness' etc. but also to any concrete noun, say 'mother'. Look at the below example in Malayalam.

(17) avaL ammy-A-y-i van-n-u
she mother-become-PST-CONJ.P come-PST-FIN
'1. She became a mother and came'(Process)
'2. She came as a mother'(Status)

The surface form *ammayAyi* is understood as 'as a mother' in status interpretation, but understood as outlined through time - 'became a mother' in process interpretation. This is similar to the ambiguity pointed out for adjectives in example 14.

Curiously, the morphological form of the verb is same and the morphology shows alleged tense markers; yet the verb inflection is non-finite. e.g. In the table 3.2, the word *niRaiya* is from the stem *niRai* 'to be filled'. As a process, it means 'as [something] gets filled' but as a status it means 'many/much'. Interestingly, the inflection 'niRaiya' shows a non-past tense inflection and it is a non-finite participial

form of the verb. The word *taLLi* is from the stem *taLLU* 'push' and it shows past tense inflection; this again is a non-finite participial inflection of the verb. How can non-finite participial inflection show tense markers and why do such markers occur in atemporal postpositions as well? These questions shall be addressed later in next chapter. For now we have just discussed that the languages exploit the same *verbiness* to create function words such as postpositions, quantifiers, complementizers, conjunctions, adjectives, adverbs.

3.5 Schemas along the dimension of interaction

Whenever an expression is uttered in a discourse, it can be understood in terms of what is the larger role this expression plays in building up the discourse. At any given time in discourse, there are two possibilities of construal available at the speaker's disposal:

1. The speaker construes that the current expression does not complete the discourse and expects more expressions that will lead him towards the nuclear/central idea that he intends to build.

2. The speaker construes that the current expression does not expect anything more and is the nuclear discourse idea itself.

For example, suppose a speaker utters the word 'The boy', there are two possibilities: either the speaker conceptualizes he is going to build a larger discourse idea, towards which goal the expression 'the boy' will eventually lead him. Here the construal would be: *The boy did what?, The boy is in what configuration?, What is the description of which the boy is a part* and so on. The second possibility is that the current expression is by itself the intended nuclear idea in the larger discourse. Suppose the discourse is a context in which somebody asks a question *Who went inside the room?* and the speaker replies *The boy*, he intends this expression as the nuclear idea itself in the discourse. Therefore the expression expects nothing more and completes the discourse. This idea can be seen in verb expressions as well. If one utters 'He went to the canteen and ate a sandwich', at the point when the speaker says the expression 'went', he intends to build the discourse further towards a larger point in discourse 'he ate a sandwich'. It is towards this nuclear idea that the current utterance 'he went' is leading to. Thus from a discourse point of view the event 'went' does not complete the discourse whereas the event 'ate' brings it to completion.

The expression which is construed as not yet completing the discourse creates some further expectations. There are two possibilities: An expectation of a *relation* is created or an expectation of a *thing* is created. For instance, when one utters the word *quickly*, it creates an expectation of some *process* in the discourse, say *walked quickly* and so on. However the utterance of the word *contemporary* prototypically creates an expectation of a *thing* that is going to be described in discourse: say *contemporary literature, contemporary culture* and so on. Of course, in English, the adjective *contemporary* can be used as a predicate where the speaker can construe the expression as bringing the discourse to completion. e.g. 'The issues are contemporary'. But for any given expression, it can be seen that the expression

can be construed to expect a thing to further build the discourse, expect a relation to further build the discourse or close the discourse to completion.

These discourse ideas of *relationship* expectation, *thing* expectation or *discourse completion* are invoked through morphological inflections of verbs in Dravidian languages. We call these discourse conceptualizations as *continuative, combinative* and *complete* schemas respectively.

3.5.1 Continuative Schema

Any entity which conceived as creating an expectation of *relation* in discourse is called 'Continuative'. Whenever an interaction happens such that a new relation is introduced as a head, the discourse continues further and in this sense these constructions are said to be in continuative schema. For example, in a structural linguistic terminology, if an adjective modifies a verb, an adverb modifies a verb, a noun is conceived as a participant of an action with some theta role, a noun is conceived as a part of some configuration through postpositions, then in all these cases some *relation* is a head of interaction. Therefore, all these kinds of interactions are said to invoke continuative schema. In a continuative schema, the target of interaction is a *relation*.

3.5.2 Combinative Schema

Any entity which conceived as creating an expectation of a *thing* in discourse is said to be in 'Combinative' schema. Whenever an interaction happens in discourse such that a new entity is introduced as a head, the discourse does not continue but anchors its description around a *thing* by combining the expressions and in this sense these constructions are said to be in *combinative* schema.

For example an attribute adjective, genitive case marker, dependent words in a noun compound, relative clause constructions are all indicative of anchoring the discourse with a *thing* as the head. Such constructions are 'combinative'. In a combinative schema, the target of interaction is a *thing*.

It can be seen that the above three schemas are highly abstract notions in discourse. There are more concrete schemas which inherit these abstract schemas, define more specific discourse conceptualizations and characterize the construction patterns observed in Dravidian languages. We discuss such schemas below.

3.5.3 Participant Schema

This is a subtype of *continuative* schema. A continuative schema in which the source of interaction is a 'thing' is said to be in participant schema. For example, nouns playing some meaningful semantic role in an event, nouns which are in some configuration described by postpositions (basically case assigned nouns) are said to be in participant schema.

Participation can be of two types: (a) Karaka relations (b) Non-karaka relations. If a noun conceptually has a direct involvement in the unfolding of a *process*, it is said to be in *karaka* schema. If it plays

some meaningful role in some configuration but not directly relevant to the process unfolding, it is said to be in *non-karaka* schema. *Karaka roles* are the notions that are already available in the traditional Sanskrit Grammar of India and therefore we are not going to reinvent the wheel and discuss these participant roles in detail. These karaka roles explain the assignment of cases in meaningful ways. Please refer to these relevant works [17], [9],[29] and other publications on *Karaka theory* to understand how these conceptual roles are different from the purely semantic roles such as *Agent, Goal, Patient etc.*

For simple understanding, we briefly point out the difference between karaka and non-karaka roles here. Look at the below example.

(18) rAman tann-uTaiya vITT-il mOhan-OTu paTam pArttukkoNTiruntAn
Ram self-GEN house-LOC Mohan-ASSOC movie was watching

'Ram was watching a movie with Mohan at his home'

In the above sentence, the action of 'watch' requires the conceptual roles *who watched, what was watched, time and spatial containers for the action* which are integral to make sense of the relation profiled by the process 'watch'. But a concept such as 'with whom did one watch' is not directly relevant to the meaningful understanding of the act 'watch'. In transaction verbs *who gave, what was given, the directed goal, the source of separation* etc are integral part of the process unfolding. Roughly speaking, these integral conceptualizations behind the participant roles of a noun in a process are what are called *karaka* roles. Other relations such as *with whom, for the sake of what, around, about* etc are not directly relevant to the process unfolding and are called non-karaka roles.

3.5.4 Associative Schema

This is a subtype of *continuative* schema. A continuative schema in which the source of interaction is a 'relation' is said to be in associative schema. For example, in structural terms, adverbs modifying the adjectives, adverbs modifying verbs, a verb modifying another verb are all associative interactions. Since adjectives, adverbs, verbs are all treated as relations in cognitive grammar framework, wherever *relation-relation* interaction occurs, we call such interactions as associative. All event-event relations in discourse are associative schema interactions. We will discuss these in detail in next chapter.

3.5.5 Qualifier Schema

This is a subtype of combinative schema. *Qualifier* is a schematic notion where a relation is conceived such that it describes a *thing (noun)* as an integral part of the unfolding of the relation in any world of discourse. What this means is that conceptually any construction unit that attributes some relation on a noun is said to invoke *qualifier* schema. In other words, the construction units which are in qualifier schema semantically expect a *thing* in the discourse to satisfy the qualifier's functional expectation.

For instance attributive adjectives, genitive case marker, relative clauses, noun quantifiers, determiners, demonstrative adjectives, appositions, attributive quotative markers etc. signify some relation and functionally describe a noun as being a part of that relation. Hence all these various grammatical units have the generic discourse function of invoking a qualifier schema. Qualifier schema attributes a relation on some noun and therefore it inherits *verb* schema.

In fact, Dravidian languages exhibit verbal morphology to indicate wherever qualifier conception is invoked in a discourse. Remember that all relations are verbs and therefore qualifier is actually indicated, in these languages, by the relative participial inflection of verbs. To appreciate this fully, look at the below examples.

(19) **oru** paiyan vantAn
One-qualifier boy come-pst-3P-male-sing
'**A** boy came/ **One** boy came'

(20) **enn-uTaiy-a** naNpan va-nt-An
I-belong-RP friend come-PST-3.M.SG
'**My** friend came'

(21) **azhak-A-n-a** pATal-ai kET-T-En
beauty-become-PST-RP song-ACC hear-PST-1.SG
'I heard a **beautiful** song'

(22) ushhh **en-R-a** oli ezhu-nt-atu
ushhh say-PST-RP sound arise-PST-3.NHUM.SG
'The sound **(quoted as)** "ushh" arose'

(23) tapAlkAran**A-n-a** suresh nERRu va-nt-An
postman-become-PST-RP Suresh yesterday come-PST-3.M.SG
'Suresh, **(who is)** the postmaster, came yesterday'

(24) amIr-in A,B,C,D **mutal-A-n-a** patan-kaL veRRi
Amir's A,B,C,D first-become-PST-RP movie-PL success
peR-R-ana
attain-PST-3.NHUM.PL
'Amir's movies **namely** A,B,C and D achieved success (in the box office)'

(25) gOkul **ezhut-iy-a** nUl mikunta varavERp-ai
Gokul write-PST-RP book great welcome-ACC
peR-R-atu
receive-PST-3.NHUM.SG
'The book **that Gokul wrote** received a great welcome'

In the above list of examples, the entities which are highlighted in boldface are in *qualifier* schema and the underlined entities are in *noun* schema. Note the morphological regularity of RP inflections of verbs in every one of these cases. Genitive case marker in example 20, attribute adjective in example 21,

relative clause in example 25 exhibit the same RP inflection in their morphology. Since the language is head-final, the attributing entity which is in *qualifier* schema always precedes the noun as shown by the boldface and underlined entities. Even relative clauses are expressed as relative participial form of verbs that precede the noun as a modifier. Appositions such as 'suresh, the postman' in example 23, quotatives such as 'the sound "ushh"' in example 22, noun modifiers such as 'movies namely' in example 24 etc. exhibit the same *relative Participial (RP)* inflection morphologically.

The example 19 shows the word *oru* 'one' which is not an RP inflection of a verb per se morphologically. But syntactically this form *oru* occurs only when it functions as an attribute adjective, but becomes *onRu* as a predicate. The idea is, even if the morphology does not show the relative participial inflection in this case, the attribute adjective form is still distinct from the predicative form. The functional generalization that we want to point out is, Dravidian languages exhibit a construction template called as *qualifier* schema whenever some relation has to be attributed on a noun in discourse.

3.5.6 Compound Schema

Compound is a subtype of *combinative* schema where the source of interaction is a 'noun'. When a noun interacts with another noun in discourse, such a conception is called as combinative schema. Noun in genitive case, every non-head noun in a multi-word noun compound expressions - are said to be in combinative schema.

This perception of one noun becoming a part of another noun or becomes a part of a relation can change how you understand the whole sentence itself. Identifying whether a noun is in combinative schema or not is extremely important to understand Tamil sentences because noun compounds are not some conventional expressions but are expressions created on the fly. The interpretation of the sentence might change based on whether the noun is perceived as combinative or not. Look at the below sentences in Tamil.

(26) makkaL kuRai tIrppu ANaiyatt-iTam muRaiyiT-a se-nR-Ay
People grievance redressal commission-LOC appeal-INF go-PST-2.SG
'You went ahead to appeal to the 'Public Grievance Redressal Committee"

(27) makkaL kuRai tIrppu ANaiyatt-iTam muRaiyiT-a se-nR-anar
People grievance redressal commission-LOC appeal-INF go-PST-3.H.PL
'People went ahead to appeal to the 'Grievance Redressal Committee"

The examples 26 and 27 are very interesting. In 26, the first word *makkaL* 'people' is interpreted as a part of the whole expression *makkaL kuRai tIrppu ANaiyam* to mean 'Public grievance redressal commission'. This is because the last verb *se-nR-Ay* 'You went' suggests that the doer of the action 'go' is some second person and not the *makkaL*. Since Tamil is a pro-drop language, the second person pronoun is dropped and the second person agent is inferred only from the verbal inflection. Once the possibility of agency is ruled out for the word *makkaL*, the reader interprets the the word as if it is

interacting with the remaining nouns and treats it a part of the multi-word expression and thus said to be in 'compound' schema.

Whereas in example 27, the same word *makkaL* is treated as if it were the doer of the action 'go' because the verb *se-nR-anar* 'they (humans) went' suggests that the agent is some third person human plural which aligns with the word 'makkaL'. In this case the noun 'makkaL' is conceived to play the conceptual role of doer/ initiator of the action and thus it is said to be in 'participant' schema.

We have discussed about noun, verb, process, status, continuative, combinative, qualifier, compound, associative and participant schemas.

3.5.7 Pronominal Schema

This is a discourse conceptualization that is an interesting special case of a *Noun* schema. Structurally, expressions that I analyze as invoking *Pronominal schema* correspond to the headless relative clauses, predicate adjectives, predicate genitives, pronouns, quantifier predicates and so on. Conceptually, *pronominal* schema is an extension of *qualifier* schema with the difference that the relation is attributed not on a specific noun in the discourse, but on the least specific / most generic *thing* in the discourse. In other words, the specificity of such a *noun* comes solely by virtue of the attribution of some relation on it and therefore the resulting construction is actually a *referring expression* in the discourse. All referring expressions in discourse instantiate *pronominal* schema. We call this conceptualization as pronominal because the qualified least specific entity in the discourse stands "in place" of some other specific *noun* and in that sense it is 'pronominal'. The following examples clarify the point.

(28) rAman vITT-iRku pO-n-An
Ram home-DAT go-PST-3.M.SG
'Ram went home'

(29) **pO-n-a(v)-an** ena-kku phone sei-t-An
go-PST-RP-PRON I-DAT phone do-PST-3.M.SG
'**The (male) one who went** made a phone call to me'

In the example 28, the noun *rAman* is a specific noun with its own semantic attributes (a male, singular human whose name is Ram) just like any other noun chair, table, phone etc. But the expression *pOnavan* highlighted bold in example 29 refers to the most generic masculine entity in discourse whose only specific identity comes by virtue of being a part of the process 'going'. Thus it becomes a referring expression to 'Ram'. The expression *ponavan - the (male) one who went* is a little awkward in English because a simple 'He' is usually preferred in English under similar contexts. But such constructions are pretty regular in Tamil, to create referring expressions in discourse. As you can see, the referring expressions are indicated by a pronominal suffix added to Relative Participial(RP) inflection of verbs i.e. pOnavan = pO(go) + n (pst marker) + a (Relative participial) + an (a human masculine gender suffix).

The same morphological strategy of creating an RP inflection and adding a pronominal suffix is exploited to create referring expressions of all levels of schematic complexities, not just simple referring expressions shown in the above examples. For instance take a look at the below examples in English.

- John remembered **reading** the book.
- **He** indeed remembered it.
- **That it is** John's favourite book is known to his friends.
- The book is **yours**.
- All **those who are interested** in the trip can join us.
- Explain me **what** you **have understood**.
- John liked **the fact that** the story **had** the most interesting climax.

All the above constructions highlighted in bold invoke *Pronominal* schema in the discourse world. The complexity of these constructions corresponds only to the schematic complexity of the referring expression, but essentially all these various syntactic units invoke the same concept. **He/She/It/They, [he/she/it/one] who/which [process/status], yours/mine/ours, the fact that [process/status], [verb] that [process/status]** - all these constructions invoke the same function of referring to some least-specific/most-generic entity. The Table 3.3 explains the interpretation. All the syntactic structures

Table 3.3 Referring expressions with varying schematic complexities

Example	**What is referred to by the construction**
saw him **going** into the room	The event - *He goes into the room* in the discourse
It went	Some non-human entity in the discourse
That he **is** my friend	The idea that some configuration is a predicate
The fact that you **came**	The idea that some event is a predicate
The book is **yours**	An entity which is a part of some configuration
what you **have told** me is true	An entity which is a part of some event

mentioned in the table which are formally different, exhibit the same common discourse property of creating a referring expression in discourse. It is precisely this property that is encoded by the morphological regularity in Dravidian languages. In all these apparently various syntactic environments the same *RP+pronominal suffix* construction occurs in Dravidian sentences.

The table 3.4 shows the same English constructions in various Dravidian languages. It can be seen that regardless of the schematic complexity, the same *Pronominalized relative clause construction - RP+pron* figures in all the cases. The highlighted words are: In Malayalam: pokunnat = pok-unn-a (going-PRES-RP) + t (non-human singular suffix); In Tamil, atu = a (distal marker) + tu (non-human singular suffix); In Telugu an-E-di = ane (say / quote-NPST-RP) + di (non-human singular suffix); In

Table 3.4 Pronominalized Relative clause construction - RP-pron

Example	**Discourse reference interpretation**
avan muRiyileykk **pOkunnat** kaNT	The event - *He goes into the room* in the discourse
atu pOnatu	Some nonhuman entity in the discourse
vaDu nA mitruDu **anEdi**	The idea that some configuration is predicated
nI vandAy **enbadu**	The idea that some event is predicated
pustaka **ninnadu**	An entity which is a part of some configuration
nI **paRanjat** satyam AN	An entity which is a part of some event

Tamil: enpatu = en-p-a (say / quote-NPST-RP) + tu (non-human singular suffix); In Kannada: ninna-du = ninna (your) + du (non-human singular suffix); In Malayalam: paRanjat = paRan-j-a (tell-PST-RP) + t(non-human singular suffix).

Even the basic pronouns are conceptually derived in the same way - a *Qualifier* relation attributed on the most generic entity. This is reflected in the morphology of their derivation. There are no basic pronouns like *he, she, it, they etc.* found in these languages. Only pronominal suffixes are there that denote the noun category and a distal, proximal or indefinite markers combine with pronominal suffixes that results in a referring expression. For example in Kannada, there are 5 grammatical categories into which any noun can belong to: singular human male, singular human female, singular non-human, plural human, plural non-human. So there are five morphemes *-anu, -aLu, -du, -aru, -vu* which correspond to these five categories. Even a simple pronoun like *He* is derived by adding either a distal marker *a+anu = avanu (distant singular human male)* or a proximal marker *i+anu = ivanu (proximal singular human male)* or an indefinite marker *yA+anu = yAvanu (which male)* - these a/i/yA markers attribute their distance/proximal/indefinite configuration on some generic entity belonging to a noun class and hence become referring expressions. Same is the derivation for other pronouns as well. This is true for other languages such as Telugu, Tamil, Malayalam as well.

Basically, through these discussions we point out that discourse and the functional conceptualizations involved therein are the basis for the peculiar morphological inflections that regularly occur in various seemingly different syntactic environments. The best way to label these morphological forms is not formal *grammatical parts of speech* but functional *construction schema labels*. Now obviously the subtypes of *qualifier* and *pronominal Schema* naturally are: **process qualifier, status qualifier, process pronoun, status pronoun**. Other schematically complex subtypes are **quote qualifier, quote pronoun** (i.e. *[noun] that [process/status]* construction is quote qualifier, *the fact that [process/status]* construction is quote pronoun). With this we come to an end of the discussion of the following schemas: *noun, verb, process, status, continuative, combinative, participant, associative, qualifier, compound, and pronominal, process qualifier, status qualifier, process pronoun, status pronoun, quote qualifier, quote pronoun*. The next chapter explores other schemas related to tense, aspect, modalities and event-event relations in discourse.

Chapter 4

Event anchoring and interactions between multiple processes

As discussed in chapter 2, tense and finiteness are separated in Dravidian languages[4]. Every complex sentence is made up of multiple non-finite verbs and only one finite verb[26]. Every one of those non-finite verbs show tense morphemes[5]. The only type of sentence where more than one finite verb can occur is reported / attributed clauses when one sentence is quoted within another sentence. It has also been suggested in literature that these so called tense morphemes in Dravidian verbs are aspect morphemes. We make the following claims about Dravidian *process* verbs based on the concepts that the process verbs invoke in the discourse.

1. Markers understood as indicative of tense in Dravidian language *process* verbs are not really tense markers. These markers are merely shift in the cognitive viewpoints of a given event. We will discuss about these viewpoints in the next section.

2. Finiteness/ Non-finiteness of Dravidian *process* verbs is a reflection of complete/ non-completion of narration of an event in discourse. The most independent and the nuclear utterance corresponds to finite inflection and all other non-nuclear/ satellite utterances invariably show non-finite inflections directly expressing the discourse event relations.

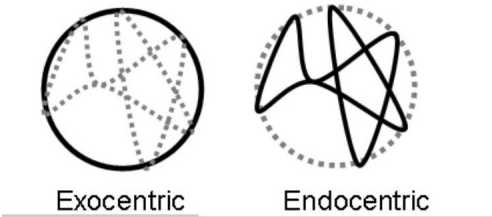
These two points are elaborated with relevant linguistic evidences in the following sections. Before venturing into a detailed discussion and implication of these two points mentioned above, let us understand what is meant by exo-centric and endo-centric viewpoint of an event denoted by a process verb.

4.1 Exo-centric and endo-centric Viewpoints of Process in Discourse

In cognitive grammar, every *process* verb unfolds a relation conceived along a temporal axis through sequential scanning[35]. This is unlike the configuration *status* verbs which are atemporally available through summary scanning[35]. The temporal unfolding does not correspond to the actual time but to how the relation is sequentially viewed by the speaker as he conceives it. Now this perspective of event unfolding can be exo-centric or endo-centric. Exo-centric viewpoint describes a situation as an integral whole, whereas the endo-centric viewpoint describes a situation in terms of its internal structure[28]. In

an exocentric viewpoint, the speaker gets a perspective as if he is outside the mental space within which the event unfolds while in an endocentric viewpoint, the speaker gets a perspective as if he is within the space where the event unfolds right before him with all the internal details. The figure 4.1 shows these two viewpoints[28]. At the level of discourse function, this translates to (i) whether the speaker thinks

Figure 4.1 Cognitive viewpoints of an event



of an event as already manifest - (hence you can view it as an integral whole)- in the discourse world and available as a knowledge for description or (ii) whether the speaker thinks of an event as manifesting now - (hence all the internal details of an event are accessible) - in the discourse world as a dynamic conception for description. Grammatically, this corresponds to perfect or non-perfect aspects of a verb. It is these manifest/non-manifest conceptualizations that are marked as past/non-past *(so called) tense* morphemes in Dravidian verbs. Let us look at the example below.

(30) AryA nayantArav-ai anb-OTu pArp-p-An
Arya Nayantara-ACC love-ASSOC see-FUT-3.M.SG
'1. Arya will look at Nayantara with love' (absolute future)
'2. Arya would look at Nayantara with love' (probable past)
'3. Arya used to look at Nayantara with love' (Habitual past)
'4. Arya looks at Nayantara with love' (General fact or Histoical present)

In the above example the process *pAr***p***pAn* 'He will see' has a morpheme **p** which is typically analysed as a future tense marker in Tamil. But it does not mean future tense per se. All it says is that, the action denoted by the verb is viewed as a new information getting unfolded in the world of discourse. Therefore the verb *pArppAn* has at least four interpretations in various discourse contexts as revealed in example 30.

To put the idea into context, let us assume that my friend asks me: "Hi, you watched the new movie, didn't you? What happened in the movie?". And I narrate a scene to him as if it is unfolding before us in the movie world (The movie that I watched sets the world of discourse in the past). Instead of narrating the event as if I am remembering the event of my past experience, I want to present the unfolding of the events in the movie as if we are viewing it in front of us. So the narration of the story would use same future marker as follows. The glosses are not given for the entire discourse but the verbs are shown in bold with nonpast morphemes 'v','p' highlighted in italics. The translation is an approximate rendering of what is expressed in the source text.

paDattil AryA nayantArAvai mikavum **virumpu*v*An**. avaLiTam kAtalai solliviTa **ninaip*p*An**. nayantAra avan arukil **varu*v*AL**. AryA nayantArAvai anbOTu **pArp*p*An.**

In the movie, Arya **would be** madly in love with Nayantara. He **would want** to propose her soon. Nayantara **would come** near him. Arya **would look** at her with love.

Look at another discourse context where a habitual past description is made.

engaL vITTin arukil oru TAktar iruntAr. avar nALtoRum kAlaiyil walking **pO*v*Ar**.

There was a Doctor living near my house. He **used to go** for a walking everyday.

As we can see, the verbs in Tamil have the same morphemes 'p' and 'v' is used to mean the habitual past action of *used to go*, a non-perfect aspect in that past discourse world. Thus the future marker is only an aspect marker and tense is not explicitly encoded. Its relation with the utterance world is not encoded by the language since the discourse context itself would activate such a temporal relation. We are not discussing the grammatical aspects such as continuous, perfect, polarities etc or modalities such as *should, must* etc. until we introduce inter-process construction schemas in later sections. For now it suffices to say that any single process can be conceived as manifest or non-manifest and that precisely is what is encoded in the language, not actual tense. These manifest and non-manifest conceptions of a process are expressed using past and non-past markers in Tamil morphology. This functional analysis that we are pointing out corresponds to the suggestions made by [3].

The same discourse idea of the aspect markers holds good for other Dravidian languages as well, but it should be remembered that the construction schema chosen could be different in different languages. For instance, to indicate the habitual past Tamil chooses a finite-verb form with the non-past marker within, whereas Telugu chooses a non-finite pronominal form of a verb with the same non-past marker within. i.e. In Tamil *vA* 'to come', inflects as *varu-v-An* 'he come-FUT-3.M.SG' to mean 'He used to come'. In Telugu *vaccu* 'to come', inflects as *vacc-E-vADu* 'come-NPSt-RP-3.M.SG.PRON' (literally, 'the one who comes / will come') to mean 'He used to come'. Hence the construction that each language symbolically chooses to pair the desired discourse function with, could vary from language to language. This is just in alignment with the theoretical assumptions of cognitive grammar where the symbolic pairing of form-function is arbitrary and different languages encode the desired function differently.

But it can be observed that all these languages use a non-past inflection to mean all the various *hypothetical result, absolute future, habitual past* etc. wherever the event is non-manifest/yet unfolding in the discourse. Unlike the other three languages, Malayalam language exhibits far more interesting morphological phenomena to bind the discourse world to the utterance world explicitly which is beyond the scope of our discussion. But the discourse basis of past / non-past marker holds good for Malayalam verbs as well. e.g. *avan varum* can mean 'he will come/ he would come/ he comes (generally)'. Thus we have two more construction schemas namely **Manifest Schema** and **Non-manifest Schema** that corresponds to a process verb in past or non-past aspect respectively.

What about the 'present tense' markers used in verbal inflections in these languages? Do they represent tense? For instance, what is the function of the present morpheme 'kinRu' in Tamil or 'unn' in Malayalam etc.? Our analysis is that the present marker indicates only a duration aspect of a non-past event and not the actual present tense. Look at the below examples:

(31) ravi rojU twaragA nidra **po-tA-Du**
ravi everyday early sleep go-FUT-3.M.SG
Everyday, Ravi **goes** to sleep early.

(32) rAman eppOzhum enn-oT inglIsh-il **samsArikk-um**
Raman always I-ASSOC English-LOC speak-FUT.FIN
Raman always **speaks** to me in English.

(33) sUriyan kizhakkil **utikkum**
Sun east-LOC rise-FUT-3P.NH.SG
The Sun **rises** in the east.

Examples 31, 32 and 33 reveal the pattern that unlike English, the future markers are used to mean the generic non-past events in a discourse, not the present marker. Using present marker in example 31 would only mean 'Everyday, Ravi is going to sleep early (nowadays / since recently)'. Similarly using the present marker in example 27 gives the interpretation 'The Sun is rising in the east (now or nowadays or in recent times etc)'. The infelicitous interpretation of the present marker in the generic present tense is indicative of the fact that it is actually a duration aspect of a non-past event at the utterance time. Such a view is further corroborated by the fact that the present markers are interchangeable with future markers in non-finite verbs but not with the past markers. The following example demonstrates the idea:

(34) cAr OTT-um-pOtu enak-ku fOn va-nt-atu
car drive-NPST.RP-time I-DAT phone come-PST-3.NH.SG
'While driving the car, I received a phone (call)'.

The expression *OTT-um-pOtu* with the future inflection *um*, can be replaced with the present inflection *kiRa* to become *OTTu-kiRa-pOtu* and still the intended meaning of the sentence remains unaffected in discourse. The perspective induced is as if you perceive the incident as it unfolds before your eyes. By interchanging the inflection with the past marker the perspective changes. i.e. *kAr OTT-I(y)-a-pOtu* produces a difference interpretation as if you shift your perspective to the present moment and looking back at the event as an existing knowledge. The fact that the present, future markers merge together as interchangeable morphemes is again indicative of the fact that the language basically encodes the past / non-past aspects only in discourse.

Further basis of this interpretation comes from the historical evolution of the present tense morpheme. It is well understood in literature that the present tense marker is a recent addition to the inventory of Dravidian morphology under the influence of Indo-Aryan languages[7, 47, 52, 48]. Originally

there were just past / non-past morphemes in Dravidian morphology. At that stage, the non-past must have included multiple discourse interpretations such as duration aspect of an event in utterance time, historic present, generic present, absolute future, hypothetical future etc. But with the emergence of a special present marker which indicates the duration aspect, these original non-past markers probably became to be reanalysed as morphological future markers.

4.2 Discourse basis of Finiteness and non-finiteness

We had mentioned in chapter 2, section 2.3 that discourse is relevant to understand the non-finite and finite verb inflections in Dravidian languages. We mentioned for instance that conjunctive participial inflection essentially encodes discourse *continuance*. We also mentioned that whenever two events interact in discourse, the discourse relation is explicitly brought into Dravidian morphology as non-finite verbal inflection. These non-finite inflections are well studied in syntax and are called as adverbial participials. Such participial inflections are found in other Indian languages as well. For instance, the Hindi sentence *vah gussA hokar bAhar calA gayA* 'He became angry and went out' shows the verb 'hokar'(become-CONJ) in conjunctive participial inflection. There are two events 'gussA honA' (to become angry) and 'jAna' (to go) which are related to each other by this adverbial participial inflection. There are other such adverbial particpial inflections which connects two clauses.

However in Dravidian languages the same conjunctive inflection occurs in sentences where there is just one clause e.g. *vantu (come-CONJ) irukkiRAn (aux-PRES-3.M.SG)* 'He has come'. This sentence shows only one clause but the main verb should necessarily be inflected as conjunctive participial form for the sentence to remain grammatical. The same condition applies for other non-finite inflections as well: infinitive participial form is mandated in modalities, concurrent participial inflection occurs in adverbs and so on. Why do the same inflection occur in all these different syntactic environments? We propose that from a discourse functional perspective these inflections can be understood and explained better. There are distinct cognitive construals underlying these participial forms. We say that, to build a discourse, the same construals are invoked in all these formally different syntactic environments. With finite number of discourse construals innumerable combinations of formally different syntactic phenomena are generated.

The question is if the building up of discourse has to be encoded by finite number of construction patterns, there must be a principled way in which hundreds of non-finite verbal forms are generated out of finite number of inflection patterns. We will discuss that principle here.

When two events interact with each other in discourse, there are two possible ways in which they can interact.

- An entity- event interaction. This happens when a noun which is an integral part of a process becomes a participant in another process.

- An event- event interaction. This happens when one process associates with another process in some meaningful way.

With these two discourse interactions, Dravidian languages are able to handle all types of complex sentences. Let us elaborate the two possibilities.

4.2.1 Using process qualifier schema to encode a range of syntactic functions

A process qualifier schema is invoked when a relative participial form modifies a noun. This one construction schema is utilized by the language in various formally syntactic environments as shown below.

1. *nAn (I) sAppiTTa (eat-PST-RP) sAppATu (food)* - the food that I ate
2. *nAn (I) sAppiTTa (eat-PST-RP) taTTu (plate)* - the plate in which I ate
3. *nI (I) sAppiTTa (eat-PST-RP) pOtu (time)* - when you ate
4. *avan (I) sAppiTTa (eat-PST-RP) aLavukku (quantity-DAT)* - as much as he ate
5. *sAppiTTa (eat-PST-RP) mAtiri (similarity)* - as if someone ate
6. *sAppiTum (eat-NPST-RP) paTi (manner)* - in such a manner that someone may eat
7. *sAppiTTa (eat-PST-RP) niRaivu (satisfaction)* - the satisfaction that (one) ate
8. *nAngaL (We.EXCL) sAppiTAta (eat-NEG-RP) kOpam* - the anger that we did not eat

The above examples point out the usage of same construction template for various syntactic interpretations. The first two examples show that the head noun modified by RP is an argument of the verb, examples 3,4,5 and 6 show that the head noun is not an argument but a quality related to the event such as manner, extent, similarity to another event and so on. However in examples 7 and 8, the head noun is neither an argument nor an attribute directly relevant to the event, but rather the result of discourse attribution.

Thus the functional generalization is: In whichever syntactic environment a *thing* which is an integral part of a relation is construed to interact with some other relation i.e. *thing* integral to one *process* interacts with another *process* as a participant, then Dravidian languages readily employ this construction.

Look at the below subordinate constructions in English from which *thing - process* interaction can be inferred.

(1) *When I went home*, it rained - 'Time' is shared between the two events as a *conceptual container - adhikaraNa*.
(2) The boy whom *I met* in the market, *lives* next door - 'The boy' is shared between two

events. 'Boy' was a *conceptualized theme - karma* in first event and is the *conceptualized doer - kartA* in second event.
(3) Rajesh *spoke* so softly that I *could not hear* it - 'Manner' of the first action is shared to describe the second action.

It can be seen in the above examples that some noun which is integral to the unfolding of the first event such as *the conceptual doer, theme, instrument, container, directed goal, manner of action, time of action, intensity / extent / manner of action* etc. are shared with the second event. Of course the conceptual role could be different in the two events i.e. a noun conceived as instrument in first event can become a doer in second event and so on. But the basic idea is that, wherever such noun from one event interacts with another event as a Participant Schema, Dravidian languages inflect the first event verb as a qualifier schema and modify the shared noun. This explains why there are no explicit discourse connectives such as 'as much as, to the extent that, when, while, during' etc. since the functional properties of these connectives are alternatively encoded by the 'qualifier noun' constructions. The table 4.1 exemplifies such constructions. Table 4.1 shows how in various clausal scenarios, wherever some noun sharing occurs

Table 4.1 Noun integral to one event becomes participant in another event

English construction	**How information is encoded in Dravidian**
When I came home, [main clause]	**come-PST-RP time**, [main clause]
[main clause] **so well that** I **could understand**	For me **to understand be-possible-RP extent** [main clause]
As long as 'Event1' [main clause]	**Event1-CONT-RP limit** [main clause]
Eversince 'Event1' [main clause]	**Event1-ASPECT-RP** time-ABL [main clause]

between two events, the languages exploit the same strategy of qualifier noun construction to achieve the intended effect. The verb in qualifier schema showing RP inflection is highlighted in bold and the noun shared is shown underlined in Table 4.1. Thus the languages use just one construction template of qualifier schema, to cover all the specific clauses which involve an entity sharing with another clause. The second possibility is not of entity sharing, but of events themselves interacting with each other in discourse. Dravidian languages basically use four conceptualizations behind event-event interactions. We call these conceptualizations as conjunctive, concurrent, conditional, infinitive schemas. These schemas correspond to the four adverbial participial inflections of a verb. We will discuss about each one of these schema and their discourse interpretations in next section in detail.

Thus all kind of complex sentences are created by just these five construction schemas Qualifier, Conjunctive, Concurrent, Conditional and Infinitive schemas. Hundreds of non-finite inflections that are seen in Dravidian morpho-syntax are actually generated as specific instances of the above construction templates. In any sentence where there are multiple events interacting with each other in discourse world, the nuclear discourse unit is expressed as the finite verb and all other satellite discourse units are expressed as non-finite verbs. This is because the non-finite verbal inflections are direct mappings to the discourse relations that they invoke. This explains why all the sentences are expressed as series of non-finite verbs and only one finite verb at the end. The reason is, there is typically only one nuclear event

to be narrated in the discourse and all other events are dependent around it as satellite events. What if a sentence describes multiple nuclear events in discourse? What happens to Dravidian sentences in such a scenario? Look at the below English sentence.

(35) As my professor had called me, I went to her cabin where two people were already in discussion with her.

Events: My professor had called me (1). I went to her cabin (2).
There, two people were already in discussion with her (3).

(1) shares the 'cause' for (2). (2) shares the 'location' with (3). (3) happens in discourse.

Now according to the discussions we have made so far, whenever some entity from one event is shared with another event, *Qualifier - Noun* construction is chosen to achieve the effect. In this case the 'cause' is shared by 'Event 1' to 'Event 2' and 'location' is shared by 'Event 2' to 'Event 3'. Hence it is perfectly possible to create a Qualifier-Noun construction. But attempting such a construction in Tamil, results in an infelicitous interpretation.

en professor ennai **kUppiTTa** kAranattAl(1), iraNTu pEr ERkenavE avaruTan **uraiyATikkoN-Tirunta** aRaikku (2) **senREn (3)**.

The entire gloss for every morpheme is not given for the above sentence. The three *process* verbs are highlighted by bold face. Of these, the first two verbs are in RP form that share 'cause' and 'location' respectively to their subsequent verbs. The underlined entity is the shared noun. While the above sentence is grammatically correct and satisfies all the functional criteria we discussed earlier, it does not convey the original intention of the English sentence. It is not unlike the scenario where the original intention is 'John is under the table', but the conveyed idea is 'John is near the table'; Some grammatical structure may have been generated but is unfaithful to what is intended to be conveyed. So what makes the Qualifier-Noun construction infelicitous in the given context?

The Tamil sentence shown above makes it look like the speaker already knows 'two people were sitting in the shared location i.e. the cabin'. Using the qualifier construction suggests as if he already knew that two people were sitting there and he went to that place. Clearly, that is not the idea suggested by the original English sentence. Only after 'going to Professor's cabin' does the student get to notice 'two people were already sitting there' therefore both are nuclear events to be introduced new in the discourse. The lesson that we learn is that Tamil's *Qualifier* schema is felicitous as a construction template to describe an event as a satellite discourse unit but infelicitous to narrate a nuclear discourse unit.

Hence a natural rendering in Tamil would be to introduce some other felicitous discourse relation relevant to the context such as: 'Saying that my professor had called me, when I went to her cabin I saw two people sitting there in discussion with her'(roughly) . Look at how the whole sentence is represented

with very different constructions in Tamil. Tamil brings a temporal relation in event 2 so that event 2 can become satellite for the description of the main/nuclear discourse idea i.e. event 3 in this case. Such a nuclear discourse unit which is independent of any other discourse unit is said to be in *Complete* schema and is marked by finite inflection. Similar arguments hold good for Telugu, Malayalam and Kannada sentences as well. Thus we conclude our discussion by saying that, Dravidian languages show morphological non-finite inflections as a construction template for discourse relations on satellite events. The languages allow only those constructions which narrate a host of satellite discourse units centred around a nuclear discourse unit. Multiple nuclear discourse units cannot be narrated in a single sentence. Either they are expressed as multiple sentences or restructured alternatively into satellite events.

4.3 Inter-process construction schemas

In the previous sections we had just listed out that there are four conceptualizations behind event-event interactions in discourse namely *Conjunctive, Concurrent, Conditional, Infinitive* schemas. We will discuss about each schema and their conceptualizations in this subsection. Every Dravidian verb takes four basic non-finite participial inflections. Each of these inflections has a specific function in discourse. Let us begin from morphology and then proceed to connect them with the discourse function that they stand for. A Telugu verb like *veLLu* 'go' has the following four non-finite inflections,

- veLLi (go-past-adverbial participial 1 inflection)
- veLLagA (go-nonpast-adverbial participial 2 inflection)
- veLLitE (go-past-adverbial participial 3 inflection)
- veLLa (go-nonpast-adverbial participial 4 inflection)

We call these basic non-finite inflections as invoking four conceptual schemas namely **Conjunctive Schema, Concurrent Schema, Conditional Schema and Infinitive Schema**. Before we talk about the conceptualizations of these schemas, we will talk about their morphological properties.

Out of the four type of inflections, the *conjunctive and conditional* inflections morphologically show past marking while *concurrent and infinitive* show non-past marking/ bare stem. This is consistent in other verbs as well: cUDu(see) inflects as 'cUsi(conjunctive) and cUste(conditional)' where the portion 'cUs' already indicates the stem for past inflection. But 'cUDagA(concurrent) and cUDa(infinitive)' both show 'cUD' which is the non-past bare verbal stem. Similarly 'rA/vaccu(to come)' inflects as 'vacci(conjunctive) and vaste(conditional)' whereas they inflect as 'rAgA(concurrent) and rA(infinitive)' that clearly shows the morphological past and non-past distinctions in these inflections. All the other Dravidian languages also exhibit these four types of inflections with similar morphological properties. Table 4.2 shows the same four inflections in four languages for the verb 'tell'.

In Tamil, the morphological forms of both the concurrent and infinitive inflections are same and ambiguous while there are distinct forms in other languages. It can be observed that in all the four languages,

Table 4.2 Four basic non-finite morphological inflections

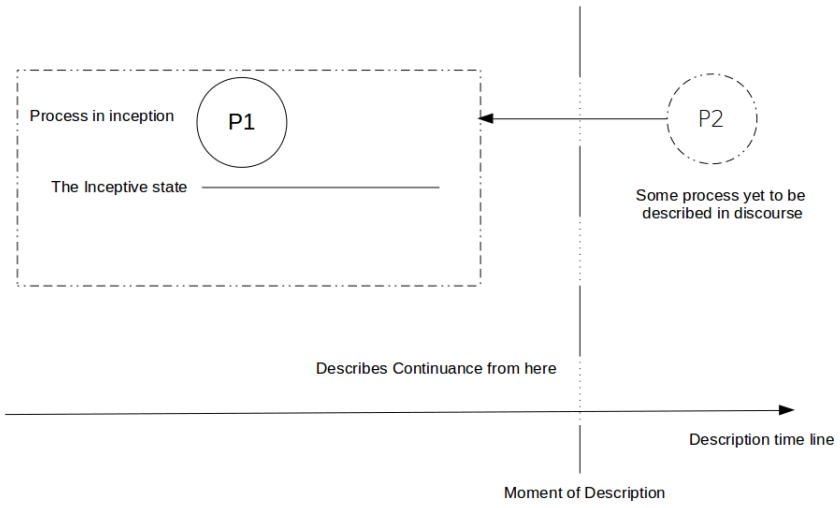
Schema Type	Kannada	Malayalam	Tamil	Telugu
Conjunctive schema	hELi	paRanju	solli	ceppi
Concurrent schema	hELuvAga	paRayave	solla	cheppagA
Conditional schema	hELidare	paRanjAl	sonnAl	cheptE
Infinitive schema	hELa/hELalu	paRayAn/paRayuvAn	solla	cheppa

morphological forms which correspond to *conjunctive and conditional* schemas show past inflection and those which correspond *Concurrent and infinitive* schema show non-past inflections. e.g. in Kannada *hELi, hELidare* show the past marker *-i* whereas *hELuvAga, hELalu* show the bare stem *hEL*. Similar is the case in other languages as well. Conceptually past / non-past forms indicate whether the *process* is in *manifest* schema or *non-manifest* schema (exocentric or endocentric). So what is the discourse function of these four schemas that compels them to take exocentric or endocentric viewpoints of a process? We discuss them one by one.

4.3.0.1 Conjunctive Schema:

A process in Conjunctive schema has a property that *its inception has already taken place in the discourse and another process awaits continuance with this process*. It means, a process P1 has already come into being in discourse and some other process P2 is expected to show continuance with it. Figure 4.2 illustrates the situation in discourse. The box in dotted lines denotes the scope of the *process* verb

Figure 4.2 Conceptualization of Conjunctive Schema



in conjunctive schema. The direction of arrow towards the conjunctive scope indicates that the current process which is conceived in conjunctive schema does not complete the discourse narration and it expects another process to build the discourse further. Since the current process is conceived as already

manifest in the discourse, we always see past morpheme in conjunctive inflection. Now there are two sub-interpretations of the above conceptualization. Either the interaction with process P2 is conceived

Figure 4.3 Event continuance in discourse

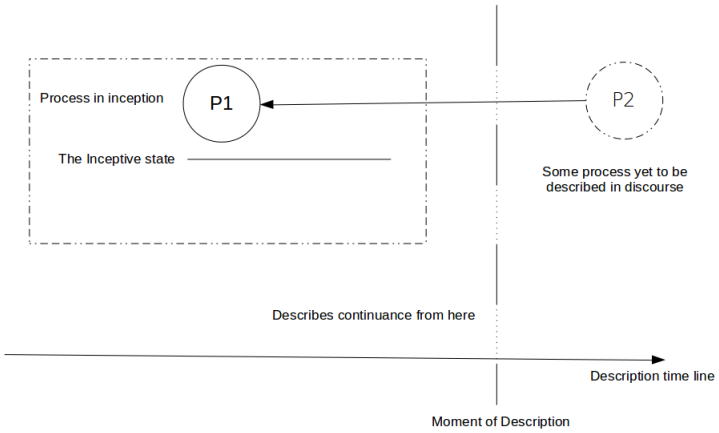
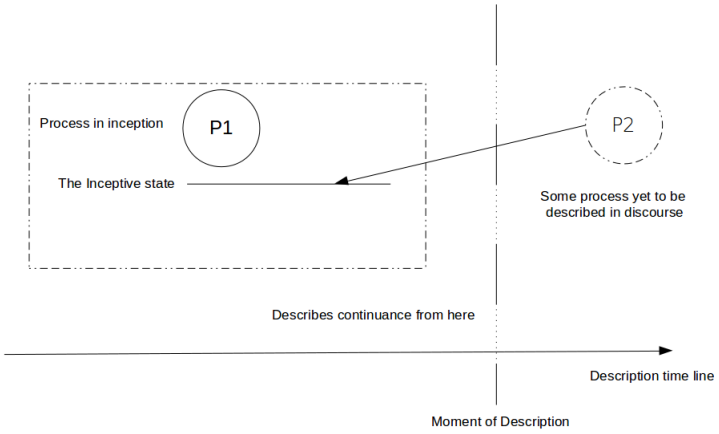


Figure 4.4 Grammatical Aspects



from the current *process* itself or with the *inception state* of the current process. The former interpretation gives rise to *event continuance* and the latter gives rise to *grammatical aspects* in discourse. These two interpretations are shown in figures 4.4 and 4.3. There are further distinctions possible even in event continuance itself, which we are not discussing in detail here. It should be elaborated separately.

The above mentioned two conceptualizations are directly encoded by the Dravidian morphology to (a) create grammatical aspects, (b) event-event continuance. The same conceptualization holds good both in affirmative and negative conjunctive schema. Look at the below constructions in English:

- I **have come** home.
- John **is going** to office.

- You **have been telling** this since morning.
- John **went** home **and drank** a cup of coffee.
- He **died jumping** into the river.

To express all the above ideas, Dravidian languages invoke the conjunctive participial inflection morphologically. We point out that this is not just a morphological process but a direct representation of the discourse conceptualization that we pointed out above.

Table 4.3 Grammatical Aspects as conjunctive schemas

Expressions in Tamil	**Literal meaning**	**Translation**
vantu irukkiREn	having come, I exist (in the inceptive state)	I have come(interpretation 2)
vantu irunten	having come, I existed (in the inceptive state)	I had come
vantu koNTu irunten	having come, having held (the i.state), I existed (in it)	I was coming
vantu iruppAn	having come, he will exist (in the i.state)	He will have/would have come
vantu viTTAL	having come, she left/let-go (the i.state)	She has come (‘already came’ interpretation)
vantu viTTu iruntAL	having come, having let go (the i.state), she will exist(in it)	she would have/will have (already) come

The table 4.3 shows how all the various grammatical aspects are expressed as conjunctive schema interpretations in discourse. The entries in the table show a series of conjunctive participial inflections and one finite inflection that give rise to interpretations of various grammatical aspects. It is a well attested fact that in verb - auxiliary verb sequences the tense and agreement(if the language marks it), is always with the auxiliary verb and not with the main verb. In formal syntactic analysis, it is considered that auxiliary verbs are syntactic heads but it is just a syntactic property and not meaningful. What we are pointing out is that the syntactic head property of the auxiliary verb is not some formal property but has a conceptual basis in terms of the **auxiliary verb interacting with the inception state of the main verb**.

This is readily apparent in the Dravidian morphology which directly encodes this idea as shown by the table above. In the example ‘he has come’, the auxiliary verb ‘has’ takes the present tense as well as the agreement with the subject ‘he’. Similarly in Tamil expression ‘avan vantu irukkiRAn’, the auxiliary verb ‘iru’ shows the present tense inflection and agreement with the subject ‘avan’. Additionally it also directly encodes that the main verb is in conjunctive schema with the auxiliary verb. It would be ungrammatical to use just the bare stem ‘* va irukkiRan’ to mean ‘he has come’. The language demands that the main verb should always be in ‘conjunctive’ inflection with the auxiliary verb.
How do we linguistically verify that auxiliary verb indeed interacts with the inception state of the main verb? The evidence again comes from the peculiarities in Dravidian morpho-syntax when it comes to negation. The scope of negation is not just possible for the action but also to its inception state. Look at the below examples.

(36) ravi **pEs-i** **irup-p-An**
ravi *speak-CONJ* *exist-FUT-3.M.SG*
Ravi would have spoken.

(37) ravi **pEs-A-mal** **irup-p-An**
ravi *speak-NEG-CONJ* *exist-FUT-3.M.SG*

*Ravi would not have spoken i.e. He would have remained silent. (Intended)

(The action alone negated. The inception state of negation is not described)

(38) ravi **pEs-A-mal** **iru-ntu** **irup-p-An**
ravi *speak-NEG-CONJ* *exist-CONJ* *exist-FUT-3.M.SG*

Ravi would not have spoken i.e. He would have remained silent. (Intended)

(The action negated. The inception state of negation that is left behind is being described)

The point that is made with the above examples is that by simply negating the *process verb* in the expression *pEsi* 'having spoken' *iruppAn* 'he will exist', as **pEsAmal* 'having not spoken' *iruppAn* 'he will exist', we do not get the intended meaning. This is because the grammaticalized verb *iruppAn* does not directly interact with the process of 'speaking' or the process of 'negating' but with the inception state of the negation. Hence an extra verb *iru* 'be / remain' is brought about to describe that inceptive state. Whereas in affirmative statements the inceptive state is inherent and not specified explicitly in morphology, in negative statements the same inceptive state comes out explicitly as an additional 'be / remain/ any functionally salient verb' which can describe the inception state.

This is true even if the auxiliary verb becomes entirely grammaticalized as a morpheme. e.g. *vantu viTTAn* 'he has come' becomes grammaticalized as *vantuTTAn* 'he has come' in spoken language. Even then, when the scope of the negation is just the main verb an extra *iru* is required to narrate the inceptive state i.e. *varAma* 'having not come' *iruntuTTAn* 'he has remained'. This explains why the same **conjunctive** participial forms are used to encode event-event **continuance** relation and also to encode **grammatical aspects**. The only difference lies in where the continuance is perceived - with the process or its inception state. The complexities associated with the scope of negation of auxiliary verb sequences as well as various interpretations that come along with it are beyond the scope of our discussion. We have just pointed out that the conjunctive inflection in the verb-auxiliary sequence is not a mere morphological marker but has an underlying conceptualization.

4.3.0.2 Concurrent Schema

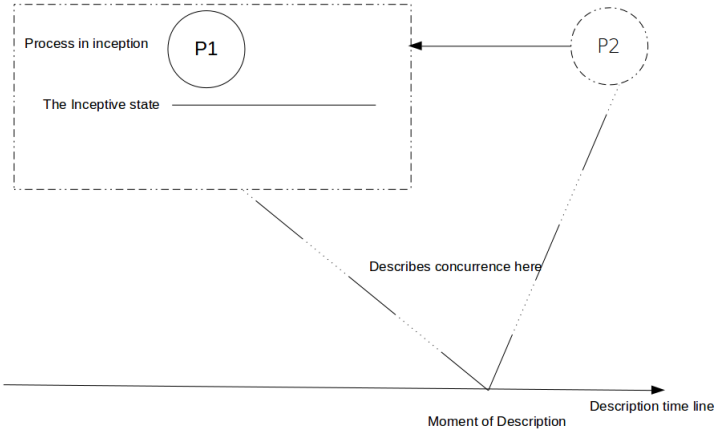
Concurrent schema is the notion that *one process is described in discourse, while holding it in perspective, another process is described in parallel/concurrent to it*. Concurrency/ simultaneity is the configuration effect produced by this conceptualization. The second type of adverbial inflection in Dravidian languages produces this effect in discourse. The following examples demonstrate this idea.

(39) bayaTaku **veLLa-gA** vAna vacc-in-di
outside *go-CONC* *rain* *come-PST-3.NH.SG*

As i **went out**, it rained.

(40) ravi kAnteenuk-ku **pO-ka** ramesh messuk-ku pO-n-An
Ravi *canteen-DAT* *go-CONC* *Ramesh* *mess-DAT* *go-PST-3.M.SG*

Figure 4.5 Concurrent Schema



While Ravi **went** to canteen, Ramesh went to mess.

The two events need not be simultaneous in real world but if conceived in the discourse world as concurrent events, they are said to invoke *concurrent* schema. If two different events which are not connected in discourse, have to be contrasted and described in one sentence, this schema is typically invoked. When one process is conceived to be concurrent with another process, the viewpoint that is invoked is endo-centric and that explains why the concurrent inflection shows the bare stem i.e. non-past inflection morphologically.

4.3.0.3 Conditional Schema

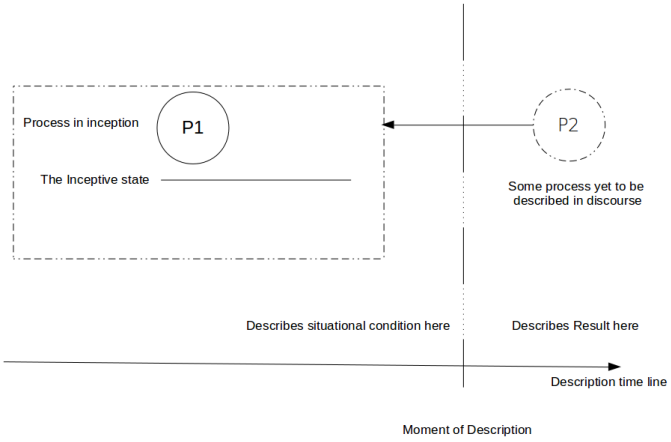
Conditional schema is the idea that *a process imposes a situational condition based on which subsequent process holds good*. All types of conditions and results in discourse, disconnected events which are conceived as if one event is the situation / condition for the description of other, hypothetical/real condition, concessions which are real/hypothetical - all these discourse ideas are expressed by conditional inflections of verbs in Dravidian languages.

(41) mazhai **pey-t-Al** payirkaL sezhikk-um
Rain shower-PST-COND crops prosper-FUT.NH
If it **rains**, crops will grow well.

(42) iNTi-ki **veLLi-te** nA amma akkaDa kUDa lEdu
home-DAT go-COND my mother there also no-exist
I went to my home and my mother was not there either.

As you can see in example 42, the first event 'I went to my home' is not a logical condition for the event 'my mother was not there'. But it is a situational condition under which I noticed the absence of my

Figure 4.6 Conditional Schema

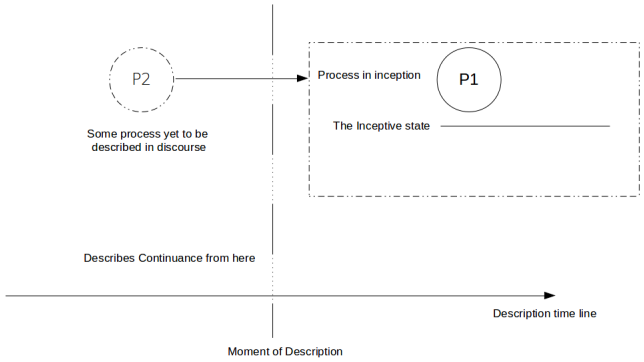


mother. Thus it invokes *conditional* schema and the languages allow conditional participial inflection in such a context.

4.3.0.4 Infinitive Schema

A process said to be in Infinitive Schema has a property that *its inception is yet to manifest in discourse as a continuance from some other process*. This is a mirror image conjunctive schema.

Figure 4.7 Infinitive Schema



While in conjunctive schema the process which is being described is already manifest and some other process is expected to show continuance with it, in infinitive schema the process which is being described is not yet manifest and is expected to show continuance with some other process. The conceptualization is shown in figure 4.7. It can be observed from the figure that at the moment of description in the time line, the process in infinitive schema is not yet manifest in discourse. This explains why always infinitive inflections always show non-past/bare stem morphologically unlike conjunctive schema which always shows past morpheme. Again there are two sub-conceptualizations possible based on whether the interaction with process P2 (a) occurs from the process P1 (b) from the

inception state of P1. Intended actions, purpose of an action exemplify the infinitive schema in the for-

Figure 4.8 Infinitive events

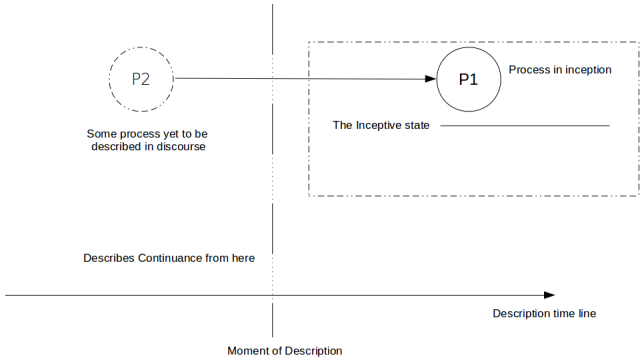
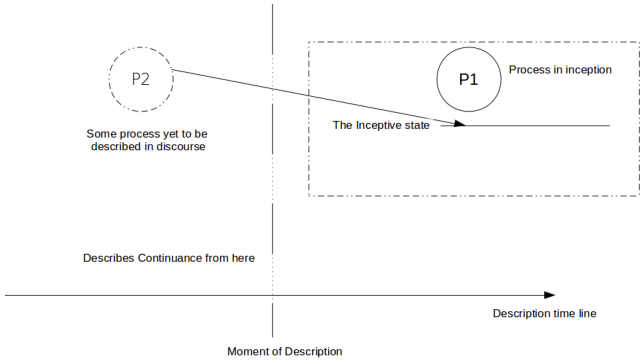


Figure 4.9 Infinitive Grammatical



mer interpretation(Process-Process interaction). Modalities such as potentiality, possibility, obligation, necessity etc. exemplify the latter grammatical interpretation(process-inception state interaction). For instance, in a sentence 'I must do' the verb 'do' is said to be in infinitive schema because while the action itself is not yet manifest in discourse, what is being described is the obligation with the inception state of the action 'do'. This is directly encoded in Dravidian morphology with a specific infinitive inflection on the main verb. Note that typically in every language, the agreement is always with the auxiliary verb because the description is not about the main verb but the interaction of the auxiliary verb with the inception state of the main verb. (Just like in the case of conjunctive schema). The table 4.4 shows how modalities are expressed as interactions between inception state of a process and another process. In Dravidian languages, two scopes of negation are possible for a *main verb - modal auxiliary sequence* which reveals that these constructions are not mere formal grammaticalized structures but have meaningful interactions. Before illustrating that point, we will show the two scope of negation in event-event interactions. Look at the below examples which show two events 'sleeping' and 'trying'.

(43) nAn **tUnk-a** muyaRci cey-t-En
I *sleep-INF* *effort* *do-PST-1.SG*

Table 4.4 Modalities expressed as infinitive schema

Expressions in Tamil	**Literal meaning**	**Translation**
solla vENTum	To say, [such a state] needed	should say
sollAmal irukka vENTum	Having not said, to remain (in such a state), needed	should not say
solla kUTatu	to say,(such a state) should not	should not say
vara muTiyum	to come, (such a state) possible	can come
varAmal irukka muTiyum	having not come, to remain (in that state), possible	cannot come
vara muTiyAtu	to come, (such a state) not-possible	cannot come

I tried **to sleep**

(44) nAn **tUnk-a** muyaRci ceyy-av-illai
I sleep-INF effort do-INF-NEG
I did not try **to sleep**

(45) nAn **tUnk-A-mal** muyaRci ceytEn
I sleep-NEG-CONJ effort do-PST-1.SG
*I tried **not to sleep**

(46) nAn **tUnk-A-mal** **irukka** muyaRci cey-t-En
I sleep-NEG-CONJ exist-INF effort do-PST-1.SG
I tried **not to sleep**

In example 45, when one attempts to say 'not to sleep' there is no morphological form that would directly negate the infinitive schema. So *tUnka* 'sleep-INF' cannot be directly negated as *tUnkAmal* or any such negative form. Attempting such a negative construction as such produces an ungrammatical sentence as shown in example 45. In order to say *not to sleep*, first you have to conceive the action 'sleep', negate it, subsequently imagine a state that stands out after negation and expect such a stage in discourse. It is this conceptualization that is encoded in Dravidian morphology. *tUnk-A-mal* 'sleep-neg-conjunctive' *irukk-a* 'remain-infinitive' *muyaRci* 'effort / attempt' *cey-t-En* 'do-PST-1.SG' i.e. 'I tried to remain in a state of having not slept'. The same logic applies to negation of a modal expression like *sAppiTa muTiyum* 'can eat'. You can either negate the 'ability' part by saying *sAppiTa muTiyAtu* 'cannot eat' or negate the 'eat' part and describe the state that remains after negation i.e. *sAppiT-A-mal* 'eat-NEG-CONJ' *irukk-a* 'remain-INF' *muTiyum* 'possible' i.e. 'I can remain in a state of having not eaten'. This is analogous to the scope of negation as seen in example 46 discussed above. Thus it is argued that modalities are not just formal semantic properties of a verb but dynamic construals showing the interactions between a process and its inception state with another process.

Basically all the various grammatical, light verb constructions, event-event interactions where an expected/intended continuance from a verb is conceptualized, the infinitive schema is invoked. It is this conceptualization that is encoded by Dravidian morphology.

4.4 Operator Schema

Now that we have discussed all the major construction schemas relevant to Dravidian syntax, we finally discuss about *Operator Schema*. An operator is any conceptualization which *takes continuative schematic entities as inputs and reinterprets/ affects the conceptual relation that these entities have with the target verb*. This means, when an operator is applied on entities that interact with a *verb* in discourse, the nature of interaction is affected/ reinterpreted. For example, particles such as *um, o, E, tAn, maTTum* etc in Tamil when added to a combinative entity, reinterpret the relation that this combinative entity already has with a verb.

- avan(He) vantAn(come-pst-agr). (He came)
- avan**A** vantAn. (Did *he* come? (Not anybody else?))
- avan**um** vantAn. (He (including others), also came)
- avan**tAn** vantAn. (Only he came. Nobody else)
- avan**E** vantAn. (He, himself, came)

At a broad level 'avan(he)' is in *Continuative* because it interacts with a process 'vA(come)' in discourse. Now by adding an operator 'E' to such an entity it modifies the nature of interaction by saying 'Only he came (no body else)'. Adding the operator 'um' to it modifies its meaning as 'He,(in addition to others) also came' etc. Because they operate on some continuative entity and changes the semantics of interaction with a verb, they are called said to invoke the operator schema. Operators cannot be added to combinative entities.

- ennuTaiya(my) pEnA(pen) - My pen
- *ennuTaiyavum(my-operator) pEnA(pen) - Pen of mine as well.
- vanna(come-pst-RP) kuTTi(child) - The child which came
- *vannavum(come-pst-RP-operator) kuTTi(child) - The child which also came
- andamaina(beauty-become-RP) mukham(face) - The beautiful face
- *andamainE(beauty-become-RP-operator) mukham - The face which is indeed beautiful

Particles which question, emphasize, list out, disjunct the options etc. add extra semantics only to a continuative entity. Trying to apply them on combinative entities results in ungrammatical constructions as shown above.

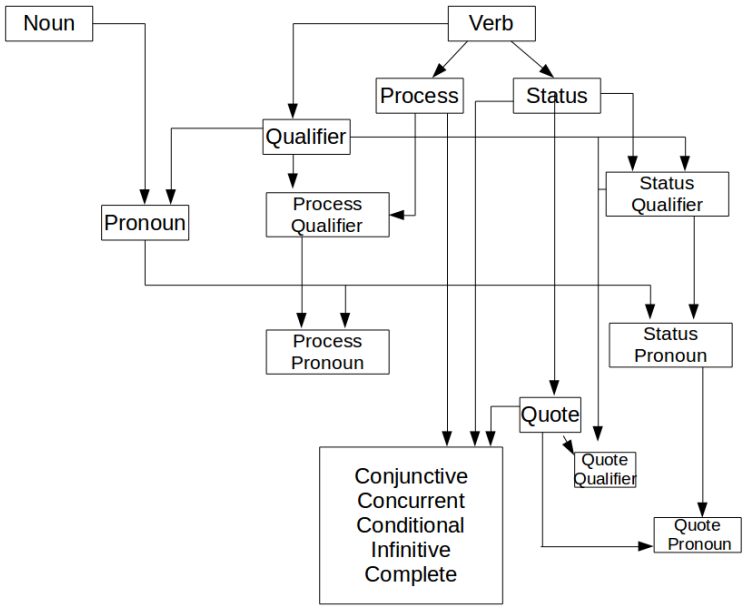


Figure 4.10 Hierarchy of Schema Derivations

4.5 Overall summary of all schemas

In this work, we have discussed all the construction schemas relevant to describe/understand the morphosyntactic structures of Dravidian languages meaningfully. We have proposed the following schemas: *Noun, Verb, Process, Status, Continuative, Combinative, Complete, Qualifier, Participant, Compound, Associative, Pronominal, Process Qualifier, Status Qualifier, Process Pronominal, Status Pronominal, Conjunctive, Concurrent, Conditional, Infinitive, Karaka, Non-karaka, Manifest, Non-manifest, Quote, Quote Conditional, Quote Qualifier, Quote Pronominal, Operator*. All the complexities that are observed in Dravidian morpho-syntax can be explained meaningfully by mapping the surface morphological forms to the above construction templates. One of the key ideas suggested by our work is that in order to describe syntax in a fully functional way, discourse conceptualizations are absolutely necessary because every sentence is not just a structural arrangement of formal grammatical units but an attempt to arrange *relations* and *things* to create a meaningful discourse. The figure 4.10 shows the larger picture of how all the major schemas are hierarchically derived from each other.

Chapter 5

CG Annotation Scheme

In chapters 3 and 4, we discussed in detail that morpho-syntactic structures in Dravidian languages can be understood as instances of construction schemas functionally. The observation we made was that discourse construals are directly encoded through regular morphological inflections in Dravidian languages. A conventional approach for Indian language parsing pipeline consists of using parts of speech tag for disambiguation during morph analysis. Thus only those morph features relevant to the POS label will be applied for morph analysis. POS tags themselves are learnt from immediate sentential context. But as we know now, learning the correct grammatical role of an expression in Tamil requires more than the immediate sentential context; a larger discourse construal has to be learnt. Therefore, instead of learning the formal parts of speech and then doing morph analysis on those units, we want to exploit these morphological regularities and see if construction schemas could be learnt better. The table 5.1 lists out the various formal grammatical units and their mapping to the construction patterns that we identified. The treebank for Tamil full parser is annotated in Computational Paninian Framework by

Table 5.1 Mapping the formal grammatical units to construction schemas

Grammatical units	**Construction schemas**
Adjectives, genitive case markers, relative clauses, appositions, quantifiers as noun modifiers, subordinate clauses as noun modifiers	Qualifier Schema (Both process and status)
Predicate adjectives, genitives as predicates, rel.clause referring expressions, predicate quantifiers, subordinate clauses signaling entity-event relations	Pronoun Schema (Both process and status)
Case markers, prepositions, subordinate and coordinate conjunctions, complementizers, adverbs	Status Associative
Discourse event-event relations, grammatical aspects, modality, complex predicates	Process associative; Subtypes are: conjunctive, concursive, conditional, infinitive
Case assigned nouns	Participant
Nouns which are part of multiword expressions	Combinative

AUKBC research centre, Chennai. The annotation framework used to develop this treebank uses the

POS, morph, chunk information alone and is not motivated from a discourse construal perspective. We need an annotation scheme that will be able to capture this.

5.1 CG annotation

A gold annotation of dependency parsing in this Construction Grammar(CG) Framework involves two stages: (a) Processing the raw text into morphosyntactic patterns for which Construction labels have to be annotated (b) Grouping these construction units into chunks and then annotating the dependency relations between these chunks. The list of Construction schemas and the dependency relations that can

Table 5.2 Tagset for Construction Schemas in Tamil

Tag Name	Construction Schema
CC	Coordinating Conjunction
ECHO	Echo Word
NN	Noun
NN_COMB	Nouns in Combinative Schema
NST	Spatio Temporal Noun
OPER	Operator
PRON	Pronoun
PSP	Postposition
PSP_PRON	Postposition in Qualifier schema
PSP_QUAL	Postposition in Qualifier schema
QT_CONC	Quotative concursive
QT_COND	Quotative conditional
QT_CONJ	Quotative conjunctive
QT_FIN	Quotative complete
QT_PRON	Quotative Pronoun

Tag Name	Construction Schema
QT_QUAL	Quotative Qualifier
RDP	Reduplication
ST_CONC	Status Concursive
ST_COND	Status Conditional
ST_CONJ	Status Conjunctive
ST_FIN	Status Complete
ST_PRON	Status Pronoun
ST_QUAL	Status Qualifier
SYM	Symbol
UNK	Unknown token
PR_CONC	Process Concursive
PR_COND	Process Conditional
PR_CONJ	Process Conjunctive
PR_FIN	Process Complete
PR_PRON	Process Pronoun
PR_QUAL	Process Qualifier

exist between them are mentioned in tables 5.2 and 5.3. A simple example is given below to illustrate the annotation scheme:

(47) timuka talaivar-um mun-nAL mutalamaiccar-um-**An-a** karuNAniti
DMK leader-also pre-day CM-also-**become-RP** Karunanidhi
nERRu seitiyALarkaL-ai santit-tu pEs-in-Ar.
yesterday reporters-acc met-conjunctive speak-PST-honorific

'Mr. Karunanidhi, the DMK leader and the ex-Chief Minister, met the reporters yesterday and spoke (with them)'

In the example 47, there is a proper noun *karunAniti* which is described by the apposition phrase *the DMK leader and the ex-Chief Minister* in English. Functionally this invokes a Qualifier schema because the noun is being described by the apposition phrase. In fact, in the above Tamil sentence this

Table 5.3 Dependency relations between construction schemas

Tag Name	Construction Schema	Tag Name	Construction Schema
ccof	Coordination conjunction	k4a	anubhava karta
conc:man	Concursive- Manner	k5	apadana karaka
conc:seq	Concursive- Event concurrence	k7	vishayadhikarana
cond	Conditional	k7a	adhikarana extended
concess	Concession	k7p	deshadhikarana
conj:gram	Conjunctive- Grammatical	k7t	kaladhikarana
conj:man	Conjunctive- Manner	nmod:stat	Status qualifier
conj:seq	Conjunctive- Event continuance	nmod	Other qualifiers
inf:gram	Infinitive- Grammatical	pof	Complex Predicate
inf:man	Infinitive- Manner	r6	Genitive case
inf:seq	Infinitive- Purpose	ras	Associative relation
k1	karta karaka	rh	Reason relation
k1s	karta samanadhikarana	main	Attachment to ROOT node
k2	karma karaka	rt	Purpose relation
k2s	karma samanadhikarana	sent_adv	Process Conditional
k3	karana karaka	status	Status functions in discourse
k4	sampradana karaka	vmod:stat	Status Associative

qualifier schema is encoded morphologically by the status verb 'Aku(become)' with RP inflection *Ana* that is highlighted above. The verb is not a process verb that describes an event, but only a status verb that describes a configuration. Therefore, in our annotation scheme the above raw text is annotated as follows:

(48) timuka talaivarum munnAL mutalamaiccarum Ana karuNAniti
NN NN NN_COMB NN ST_QUAL NNP
nERRu seitiyALarkaLai santittu pEsinAr.
NST NN PR_CONJ PR_FIN

As a first stage, the given raw text is split into construction patterns like in 48 (note that we have split the unit *Ana*, which is usually formally analysed as an adjectivializer, from the raw text since it invokes a construction pattern of qualifier). Most of the tokens in Tamil are already directly mappable to construction schemas in the above example. The labels are shown below each token.

In this way all the sentences are processed such that the tokens resulting after this preprocessing will be construction units that are ready to be labelled. The comprehensive list of construction labels and their meanings are shown in Table 5.2. After this labelling is done, the labeled construction units are chunked based on whether the consecutive construction units are *usage based syntactic freezes* or not. For example *multi-word expression(MWEs), numeric quantifiers modifying a noun, demonstrative adjectives modifying a noun* are the three instances where based on usage in Tamil, the consecutive

construction labels can be safely grouped in a single chunk. As an example the sample labelled sentence shown in 48 is chunked in the following manner:

(49) (timuka talaivarum) (munnAL mutalamaiccarum) (Ana)
(NN NN) (NN_COMB NN) (ST_QUAL)
(karuNAniti) (nERRu)) (seitiyALarkaLai) (santittu) (pEsinAr).
(NNP) (NST) (NN) (PR_CONJ) (PR_FIN)

In the above example 49, two units *timuka* and *talaivarum* are grouped together as one chunk because it forms a noun-compound MWE which is a syntactic freeze as per its usage. i.e. You can chunk them together and only the head of the chunk is going to participate in the dependency relationship. After chunking, the final dependency analysis is shown in 5.1. It can be seen that the above dependency anal-

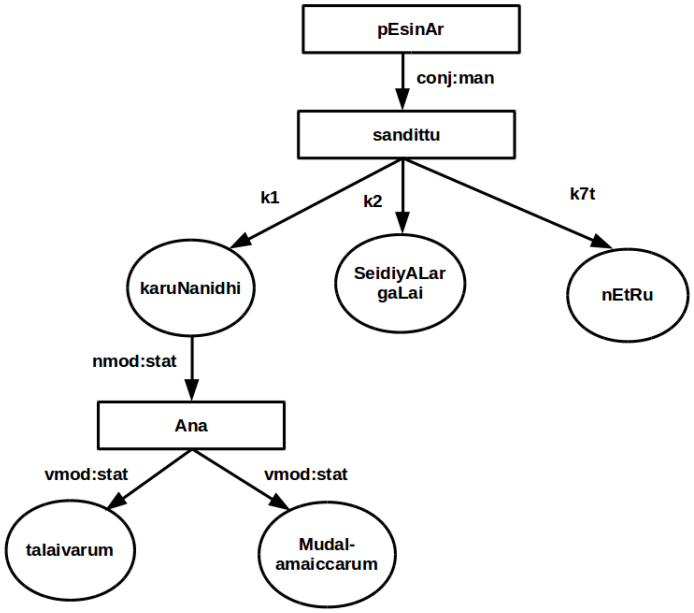


Figure 5.1 Parsed Output

ysis uses the *karaka* relations such as k1, k2, k7 as discussed in Computation Paninian Grammar(CPG) that has been successfully applied for Indian languages[11]. This is because the *karaka relations* such as *karta, karma, karana* etc. are already syntactico-semantic concepts that describe the meaningful role that a noun is conceived to play in an action denoted by a verb. These relations are well understood from traditional Sanskrit grammar and are extremely suitable for analysing relatively free-word ordered languages like Indian languages. Roughly we can say that the *karaka* roles are analogous to theta roles with the difference that these are not purely semantic from the information point of view but 'syntactico-semantic'. In that sense *karaka* roles are different from theta/ semantic roles. Refer to [11] and [9] to understand these more. In other words *karaka* theory provides mapping between form (vib-

hakti markers) and construed meaning (karaka relations) which exactly is what Construction schemas are about. Hence we are not going to reinvent the wheel but rather use these traditional notions in the general Construction Grammar formulation. As a part of the larger picture, these karaka roles exemplify the 'participant' interaction that is shown in figure 3.2. In what way is the proposed Construction Grammar annotation different from the Computational Paninian Grammar annotation? The answer can be summarized as following:

- We use all the *karaka* relations as used in CPG as these are 'construals' of how a noun can interact meaningfully with a verb.
- Additionally, we treat the function words that can occur in a language as *status* verbs i.e. postpositions, adjectives, adverbs, complementizers etc. Instead of analyzing them as formal parts of speech, we treat them as relations which describe atemporal configurations and to that extent they are all verbs in *status* interpretation. (Refer to section Chapter 2 and [35] to understand this better). In fact, Tamil uses verbal morphology to derive these function words.
- We also handle various other meaningful discourse construals such as *Qualifier, Pronominal, Conjunctive, Concurrent, Condition and Infinitive* schemas which are necessary to understand the peculiarities in Dravidian syntax. Refer to [5, 4, 38, 26, 27, 22] to understand some of the syntactic issues in Dravidian languages. Basically these discourse construals are included in our annotation framework as we hypothesize that they are inevitable to fully characterize Dravidian syntax meaningfully. Furthermore, we also hypothesize that these discourse construals are directly machine-learnable by way of morphological regularities in these languages.
- We also handle 'combinative' interaction in figure 3.2 by means of chunking multi-word expressions, noun compounds as one chunk.
- In short, we are treating the *karaka* relations as one of the four types of discourse interactions that can happen between *things* and *relations* in the speaker's discourse world. Other three types of interactions are also captured though construction labels and through chunking in our scheme.

In chapter 4 we saw that every Dravidian verb has four basic non-finite inflections that corresponds to certain discourse construals namely *conjunctive, concurrent, conditional, infinitive*. It is these construals that are skilfully exploited by Tamil to create subordinate clauses, grammatical aspects, modalities etc. by way of creating a series of non-finite verbs ending with one finite-verb. The bottom line is: formally different syntactic phenomena such as subordinate clause formation, grammatical aspects, modalities are constructed using the same above four construals. Again learning these construals instead of their formal properties is straightforward and suitable for these languages, because these discourse level construals are directly marked in morphology. For example in Tamil, *conjunctive* inflection of a verb occurs in clauses showing discourse sequence as well as in grammatical aspects. The *infinitive* inflection occurs in clauses showing goal/purpose as well as to describe moods and modalities. The *concurrent* inflection

occurs in clauses conceiving concurrence as well as in describing manner adverbs. The *conditional* inflection occurs in conditional clauses, situated descriptions, in complement clauses which are based on some condition etc. The following examples serve to illustrate the point.

(50) rAman vITTuk-ku **pO-y** paTam pArt-t-An
Raman home-dat **go-conjunctive** *movie see-pst-agr*
'Ram went home and watched a movie'

(51) rAman munp-E paTatt-ai **pArt-tu** iruk-kiR-An
Raman before-only movie-acc **see-conjunctive** *exist-pres-agr*
'Ram has already seen the movie'

Note that the same conjunctive form of a verb is morphologically used by the language to indicate both clausal sequences and grammatical aspects. Usually, in formal analyses this is explained through grammaticalization theory. But in construction grammar, it is explained as the cognitive construal that underlies the conjuctive participial inflections. (Refer to [28] for cognitive operations underlying conjunctive inflection and how it relates to grammatical aspect). Therefore in this annotation scheme, we treat every one of these non-finite verbal inflections as separate construction schemas and the dependency parser has to learn whether the relation that holds between them is event relation or merely grammatical and so on.

Thus with this understanding of the concepts that we have discussed, we formulated the construction schema labels and the dependency labels. The complete set of Construction labels in our annotation scheme is listed in Table 5.2. The complete set of dependency relations that can exist between these construction schemas are listed in Table 5.3. To understand the *karaka* related concepts such as *karta samanadhikarana, deshadhikarana* etc. please refer to the dependency annotation guidelines[10]. Other concepts such as conjunctive, concurrent, conditional, infinitive and their subfunctions namely grammatical/ manner/ event relations in discourse are already discussed. Extended adhikarana (k7a) is a *karaka* construal that is specific to Tamil. Volitionality conceived upon the giver or receiver in a transaction event is given the label k7a.

5.2 Experiments and Results

To verify if the interactions between the proposed conceptual schemas are true and suitable for learning dependencies, we collected a set of 935 sentences for our annotation experiments. Out of these, 354 sentences are taken from newspaper data that we crawled from various online newspapers in Tamil. The crawled data had a total of approximately 5 lakh sentences(5,17,421 sentences). This crawled corpus is raw, uncleaned and unprocessed. The remaining 581 sentences are taken from a portion of the gold standard annotated training data for Tamil full parser, which is available as part of ILMT consortium[1]. The gold annotation[2] of the 581 sentences was available on CPG framework. These sentences are from

[1]Indian Language Machine Translation Project funded by DIT, Government of India
[2]The gold annotation was carried out by AU-KBC Research Centre, Chennai

various domains such as tourism, health, religion etc. While the former 354 sentences are general newspaper sentences that could be on any topic, the latter 581 sentences are collected from domain specific data and therefore the type of sentences, length of sentences could differ.

We annotated all the 935 sentences with the proposed tags and dependencies. We used MALT parser[3] for the purpose of our experiments. The parser settings are of Ambati et al.[2]. The features given to the MALT parser are the stem and morphological inflections. Experiments were conducted for the following test scenarios to verify the performance of the parser.

1. Varying the sources of data (AUKBC Vs Newspaper data)
2. Varying the size of the total data (training and testing together) for the given source (354, 530, 935)
3. Varying the gold annotation scheme (existing CPG annotation Vs our proposed annotation)
4. Varying the granularity of the CG label (Granular - nmod:part, nmod:stat, nmod:act, nmod:fact Vs NonGranular - nmod, nmod:stat)
5. Combining both the data and verifying the performance with the proposed annotation

There are twelve test cases and in each case a five fold validation was performed. Two sample test case results are shown in table 5.4 and 5.5. Table 5.4 shows the results when the annotation type was Construction Grammar with more granularity (CG granular), the total number of sentences including training and testing is 354, while the data source was IIIT newspaper data. Table 5.5 shows the results for the same test conditions but with the granularity of CG annotation labels reduced. One can observe that

Table 5.4 CG granular, 354 sentences, IIIT data

IT	TR	TS	ANC-TR	ANC-TS	MNC-TR	MNC-TS	LAS % *
1	283	71	10.2	11.9	32	30	65.04
2	283	71	10.6	10. 4	30	32	67.15
3	283	71	10.6	10.4	32	29	69.42
4	283	71	10.6	10.3	32	27	76.38
5	284	70	10.7	9.8	32	26	80.14

the LAS consistently improves or at the least remains unchanged when the granularity of the labels to be learnt is reduced. It is this reduced annotation labels that we have presented finally in our dependency labels mentioned in table 5.3 under section 5.1.

It can also be seen from both the tables that when the training data has larger number of chunks per sentence and the testing data has lesser number of chunks per sentence, the LAS of the system is more as expected. The LAS score is also affected if the testing data has relatively longer sentences in comparison to training. It can be noted that the sentences in newspaper corpus that we collected at IIIT

[3]http://www.maltparser.org/download.html

Table 5.5 CG non-granular, 354 sentences, IIIT data

IT	TR	TS	ANC-TR	ANC-TS	MNC-TR	MNC-TS	LAS % *
1	283	71	10.2	11.9	32	30	70.17
2	283	71	10.6	10. 4	30	32	67.34
3	283	71	10.6	10.4	32	29	70.37
4	283	71	10.6	10.3	32	27	76.62
5	284	70	10.7	9.8	32	26	80.14

IT: Iteration number, TR: Training size, TS: Testing Size, ANC-TR: Average number of chunks per sentence in training data, ANC-TS: Average number of chunks per sentence in testing data, MNC-TR: Maximum number of chunks encountered in training, MNC-TS: Maximum number of chunks encountered in testing, LAS: Labelled attachment Score

are quite long with an average of 10.6 chunks per sentence. The average LAS scores for 5 iterations from the above tables are 71.63% and 72.93% respectively.

Similarly, 5 fold validation was done for other 10 test cases as well. Finally, in table 5.6 below, we show the average LAS score from various relevant testing scenarios. In every one of these test scenarios the result reported in table 5.6 is the average of the five iterations. Since we took only 581 sentences from AUKBC data(CPG annotation) and annotated those sentences in our proposed CG framework, the parser evaluation can be compared between the two annotation schemes for only this amount of data. Therefore, in table 5.6 shown below, the rows 5,6,7 report only the parsing accuracy on our proposed annotation scheme. CPG annotation accuracies could not be reported for these scenarios.

Table 5.6 Overall parser accuracy test cases

Ann. type	Data size	Data type	Avg. No. of chunks in a sentence	Avg. LAS
CPG	176	AUKBC	5.1	56.02%
CG	176	AUKBC	5.8	72.98%
CPG	581	AUKBC	5.2	60.57%
CG	581	AUKBC	5.9	82.24%
CG	354	Crawled data	10.6	72.93%
CG	354+176	(Crawled+AUKBC) data	9.0	74.96%
CG	354+581	(Crawled+AUKBC) data	7.7	82.21%

The annotation type *CG* refers to the proposed Construction Grammar framework and *CPG* refers to the Computation Paninian Grammar framework. The average number of chunks in a sentence mentioned in the table is the length of a sentence in terms of number of chunks in it. As it can be seen, the number of chunks are not the same between the two annotation schemes. In fact since we split a token wherever status interpretations of verbs could be made, our annotation scheme almost always has more number of chunks in a sentence than the CPG annotation. For instance, note that for the data size of 176 sentences, the average number of chunks in CPG is 5.1 whereas in the proposed CG scheme it is 5.8. The average length of sentences is more in the crawled newspaper data than the AUKBC data. The last three rows show that as the average number of chunks in a sentence in the corpus decreases, the accuracy increases which is only expected because there will be lesser long distance dependencies. Every row that is shown in the result is the average of five fold validation performed on the data. For every given data size we

chose 80% for training and 20% for testing. The average LAS accuracy of the parser on a total of 935 sentences is 82.21%.

5.2.1 Partial evaluation results of Dependency labels and attachment

Technically the LAS scores are themselves not directly comparable because the annotation labels and linguistic scheme are different. Therefore, we will show the partial evaluation results of dependency labels which reveal as to how within a given linguistic framework various labels are learnt consistently. The precision and recall of dependency labels and attachment taken together for a training size of 176 sentences annotated using the proposed method is reported in Table 5.7. Table 5.8 shows the precision and recall of dependency relation + attachment for the training data size of 176 using the CPG annotation. Similarly, 5.9 and 5.10 report the precision and recall of dependency labels + attachment for the

Table 5.7 Precision and Recall of DepRel and Attachment; 176 sentences; CG annotation

Deprel	Gold	Correct	System	Recall %	Precision %	Deprel	Gold	Correct	System	Recall %	Precision %
adv	8	6	7	75.00	85.71	k7	13	10	14	76.92	71.43
ccof	3	0	0	0.00	NaN	k7a	1	0	0	0.00	NaN
conj:gram	8	7	7	87.50	100.00	k7p	2	0	0	0.00	NaN
conj:man	1	0	0	0.00	NaN	k7t	1	0	3	0.00	0.00
inf:gram	9	8	8	88.89	100.00	main	36	36	36	100.00	100.00
k1	33	28	51	84.85	54.90	nmod:stat	5	5	5	100.00	100.00
k1s	2	2	2	100.00	100.00	pof	5	1	1	20.00	100.00
k2	12	5	6	41.67	83.33	r6	13	10	11	76.92	90.91
k2s	1	1	1	100.00	100.00	rsp	1	0	0	0.00	NaN
k3	2	0	0	0.00	NaN	sent adv	3	1	1	33.33	100.00
k4	6	1	1	16.67	100.00	status	2	2	2	100.00	100.00
k5	2	0	0	0.00	NaN	vmod:stat	21	20	36	95.24	55.56

Table 5.8 Precision and Recall of DepRel and Attachment; 176 sentences; CPG annotation

Deprel	Gold	Correct	System	Recall %	Precision %	Deprel	Gold	Correct	System	Recall %	Precision %
adv	12	7	14	58.33	50.00	k7p	5	2	9	40.00	22.22
ccof	13	9	12	69.23	75.00	k7t	5	0	2	0.00	0.00
jjmod	1	0	0	0.00	NaN	lwg__psp	2	2	3	100.00	66.67
k1	29	21	47	72.41	44.68	main	35	34	35	97.14	97.14
k1s	2	0	1	0.00	0.00	nmod	5	0	0	0.00	NaN
k2	22	10	16	45.45	62.50	r6	11	7	10	63.64	70.00
k3	1	0	0	0.00	NaN	rh	1	0	0	0.00	NaN
k4	5	2	2	40.00	100.00	rsp	1	0	0	0.00	NaN
k7	5	0	3	0.00	0.00	sent- adv	1	0	0	0.00	NaN
k7a	1	0	0	0.00	NaN	undef	3	0	0	0.00	NaN
						vmod	3	2	9	66.67	22.22

training size of 581 sentences, annotated using CG framework and CPG framework respectively. The dependency label counts are not readily comparable even for the same label because the chunks and the relationships that exist between them are different in the two frameworks. For instance, in CG output shown in Table 5.9 the gold count of vmod:stat is 90 when the training size was 581 sentences. But in CPG output shown in Table 5.10, the gold count of vmod is just 18. This discrepancy can be understood if we notice that the ccof count in the CPG annotation is 43 but in CG it is just 13. i.e. what are analyzed as conjunction of two entities in CPG are analyzed as list of entities associating with a status verb in CG. Conjunctions are actually done by particles such as 'um' whose syntactic behaviour is different from a proper conjunction like 'and' in English. It semantics is similar to MO particle of Japanese well

Table 5.9 Precision and Recall of DepRel and Attachment; 581 sentences; CG annotation

Deprel	Gold	Correct	System	Recall %	Precision %	Deprel	Gold	Correct	System	Recall %	Precision %
adv	23	18	19	78.26	94.74	k5	2	2	3	100.00	66.67
ccof	13	12	13	92.31	92.31	k7	49	39	47	79.59	82.98
conj:gram	28	26	29	92.86	89.66	k7a	2	0	0	0.00	NaN
conj:man	7	5	9	71.43	55.56	k7p	9	4	5	44.44	80.00
conj:seq	3	0	0	0.00	NaN	k7t	9	2	7	22.22	28.57
inf:gram	37	36	37	97.30	97.30	main	115	113	116	98.26	97.41
inf:man	1	0	1	0.00	0.00	nmod:stat	16	14	14	87.50	100.00
inf:seq	3	0	0	0.00	NaN	pof	18	13	18	72.22	72.22
k1	110	98	140	89.09	70.00	r6	29	25	25	86.21	100.00
k1s	16	14	15	87.50	93.33	ras	0	0	4	NaN	0.00
k2	32	20	24	62.50	83.33	rsp	1	0	0	0.00	NaN
k2s	1	0	0	0.00	NaN	rt	1	0	0	0.00	NaN
k3	9	9	9	100.00	100.00	sent adv	11	4	4	36.36	100.00
k4	17	13	14	76.47	92.86	status	11	10	11	90.91	90.91
k4a	0	0	2	NaN	0.00	vmod:stat	90	83	99	92.22	83.84

Table 5.10 Precision and Recall of DepRel and Attachment; 581 sentences; CPG annotation

Deprel	Gold	Correct	System	Recall %	Precision %	Deprel	Gold	Correct	System	Recall %	Precision %
adv	40	33	46	82.50	71.74	k7t	17	5	9	29.41	55.56
ccof	43	32	48	74.42	66.67	lwg__psp	10	4	4	40.00	100.00
jjmod	2	0	0	0.00	NaN	main	116	114	116	98.28	98.28
k1	104	77	154	74.04	50.00	nmod	13	1	5	7.69	20.00
k1s	15	2	3	13.33	66.67	nmod__relc	3	0	3	0.00	0.00
k2	63	22	59	34.92	37.29	r6	32	24	28	75.00	85.71
k2g	4	2	2	50.00	100.00	ras- k1	2	0	0	0.00	NaN
k2p	1	0	0	0.00	NaN	ras- k2	2	0	0	0.00	NaN
k2s	1	0	0	0.00	NaN	ras- neg	2	0	0	0.00	NaN
k3	6	2	2	33.33	100.00	rd	2	0	0	0.00	NaN
k4	18	12	24	66.67	50.00	rh	4	1	2	25.00	50.00
k5	1	1	1	100.00	100.00	rsp	4	0	0	0.00	NaN
k7	20	6	8	30.00	75.00	sent- adv	4	2	5	50.00	40.00
k7a	1	0	0	0.00	NaN	undef	8	0	0	0.00	NaN
k7p	27	22	39	81.48	56.41	vmod	18	15	25	83.33	60.00

discussed in literature[50]. That explains why there is more vmod:status in CG parser output. Overall, it can be seen that the dependency label + attachment precision and recall are comparatively better in our proposed CG annotation consistently.

5.2.2 Reasons for better learning

There are two reasons for better learning of dprel relations and overall LAS accuracy.

- Our dependency relations are directly learnable from morphological inflections
- The form- function pairing analysis is able to generalize the peculiar constructions that occur in Tamil syntax better.

The dependency labels that we are using in our annotation scheme are not coarser but fine grained. For instance, what is just a 'vmod' in CPG annotation is annotated with fine-grained labels such as 'conj:gram', 'conj:man', 'conj:seq', 'conc:man', 'conc:seq' etc. Yet these labels are learnt better because these labels are directly inferred from the morphological inflections given as a feature to MALT parser. The fact that construction analysis performs better across 5 fold experiments shows that most of the functional properties of the language are directly encoded in morphology.

By mapping the morphological features to meaningful construals, we are able to explain the peculiar morpho-syntactic constructions in Tamil, which are otherwise difficult or impossible to characterize. Take for example the example 5 that we discussed in subsection 2.6 from chapter 2. We repeat the same example for illustration here.

(52) nINT-a-t-um kaLaippu mikk-a-t-um-An-a
elongate-RP-PronSuffix-also tiredness exceed-RP-PronSuffix-also-became-RP
payaNam toTar-nt-atu
journey continue-pst-agr

'The long and arduous journey continued'

As it can be seen there are no pure adjectives in Tamil and therefore an expression like *long and arduous* is rendered by a complex expression *nINTatum kaLaippu mikkatumAna* which is conventionally chunked as below:

1. ((nINTatum NN)) - Noun chunk
2. ((kaLaippu NN)) - Noun chunk
3. ((mikkatumAna JJ) (payaNam NN)) - Noun chunk
4. ((toTarntatu VB)) - Verb chunk

We showed already that the most likely dependency tree that the parser will end up learning with this configuration would be as shown in figure 5.2.

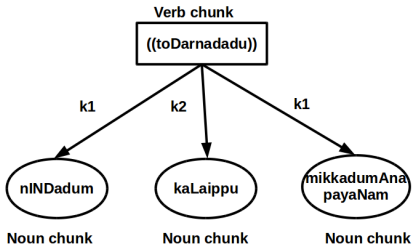


Figure 5.2 Wrong dependency that is likely to be learnt

In construction grammar approach, this problem is easily handled because we map the word 'nINTatum' as *Status Pronominal Schema*, 'kaLaippu' as *Noun Schema*, 'mikkatum' as *Status Pronominal Schema*, 'Ana' as *Status Qualifier Schema*. Thus the three words in the raw text are treated as four construction units that map to their meaningful 'construals' and now the dependency relations are established between them as shown in figure 5.3. It might look that after all the final dependency relations are the same as in CPG, so what is the CG contribution here? Recognizing 'Ana' not as a formal adjectivializer suffix but a *Status Qualifier* schema that can take 'configurations' as its arguments allows us to chunk the tokens differently and build the dependency tree shown in figure 5.3.

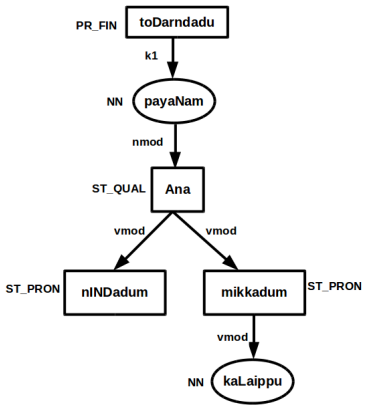


Figure 5.3 Dependency analysis according to CG framework

This and many other expressions where the language uses its morphological properties to create 'verbiness' are very well captured by following a construction based approach. In the next section, we will discuss about what kinds of errors commonly occur in parser output, analyse those error scenarios and what improvements can be introduced to handle those errors.

5.3 Error Analysis

We performed a five fold validation of parser output for varying training sizes ranging from 354 to 935 sentences and checked what are the most frequent errors that are encountered by the parser trained with the proposed CG framework. The table 5.11 shows the top five frequent errors across different folds of iteration when the training data size was 935 sentences.

Fold 1			**Fold 2**			**Fold 3**			**Fold 4**			**Fold 5**		
Gold	System	Frequency	Gold	System	Frequency	Gold	System	Frequency	Gold	System	Fequency	Gold	System	Frequency
k1	k2	11	k1	k2	16	k1	k2	11	k1	k2	20	k1	k7t	13
k1	pof	11	k1	pof	11	pof	k1	8	k7	k7p	12	k7	k7p	9
k7	k7p	6	pof	k2	6	k7	k7t	6	k1	nmod	11	k1	pof	6
conj:seq	conj:man	6	k4	rt	5	k1	pof	6	pof	k1	10	conj:man	conj:seq	6
k7	k7t	5	k7	k7p	4	vmod:stat	k4	5	k1	vmod	9	k7	k7t	5

Table 5.11 Five frequent drel errors in different folds of iteration for 935 sentences Training Data

The most frequent errors that are observed in the parser output are as follows:

1. k1 is misidentified as k2 - karta(*agent* roughly speaking) is misidentified as karma(*patient* roughly speaking)
2. pof (Complex predicate) is misidentified as k1 (doer)
3. k7p (spatial location) is misidentified as k7(general locative)
4. k4 (goal) instead of rt (purpose)

5. conj:gram (grammatical) instead of conj:man (manner)

The most frequent error is: the gold label is k1 while the system identifies it as k2 and vice versa. This usually happens because it is quite common in Tamil that the accusative case is not morphologically marked on a noun and the language is relatively free word ordered. The following sentence is taken from one of the test file in which the parser misidentified k2 as k1. It illustrates the point.

(53) vinAyaka caturttik-ku **paTam** veLiyiTuv-a-tu kamal-in tiTTam
Vinayaka chaturthi-DAT **movie** *release-RP-PRON Kamal's plan*
'It is Kamal's plan to release **the movie** during Vinayaka Chaturthi(a religious festival)'

In the above example, the highlighted word 'paTam' does not morphologically show the accusative marker. Because the parser is trained on morphological features and 'k1' and 'k2' are potential candidates in the similar context, these two tags are the most misidentified. The confusion between 'pof' and 'k1' is only expected because the noun which is a part of a complex predicate can be easily confused as a participant of the verb. The other two frequent errors are due to the granularity of the dependency relation that should be identified. *k7p* is more granular than *k7* and therefore difficult to learn. Same is true for *conj:gram* and *conj:man*. The other interesting candidate is 'rt(purpose)' being misidentified with 'k4(goal)'. Because conceptually purpose and goal are metaphorically directed towards some object, the same fourth case marker is used in Tamil to denote these two functions. Hence it is difficult to tell apart one from another. Since these error scenarios typically involve making distinctions between granular relations, with more training data these can be learnt better.

Chapter 6

A full parser system

Based on the above theoretical ideas and MALT parser experiments, we decided to develop a full parser pipeline that exploits all the theoretical notions of construction grammar that we described in chapters 3 and 4. Through our long theoretical discussions so far, we essentially arrive at a few important generalizations listed below:

- Grammar is entirely symbolic i.e. form-function pairings
- There is a continuum from lexicon to syntax. i.e. distinctions between lexicon, morphology, syntax, discourse are just differences in schematic complexities of the same semiotic system.
- Discourse concepts are essential for a fully functional characterization of syntax. Taking a sentence in isolation from its discourse context and performing *karaka* analysis alone may not be sufficient to arrive at an appropriate syntactico-semantic representation of a sentence.
- These discourse concepts are directly available as systematic morphological inflections in Dravidian languages.

How to exploit these generalizations to make policy decisions about full-parser pipeline? We adopted the following approach. We treat every raw text as arrangements of symbols of various schematic complexities. Now this arrangement of schematic complexities could be achieved in different ways in different languages. Since Tamil is an agglutinative language, this arrangement of schematic complexities is available straightforwardly as arrangement of morphological complexities. Hence given a raw Tamil text, we just split any given token (could be made up of multiple words with complex sandhi, it doesn't matter) into portions of lexicons and their inflections. Hence if an expression such as *nAykaLenkirukkinRana* 'Where are the dogs?' is one token, we want to identify that it is made up of the following symbols: 'nAy, kaL, enk, iruk, kinR, ana'. Of these *nAy, enk, iruk* are less schematic symbols i.e. lexicon and other symbols such as *kaL, kinR, ana* are schematically more complex i.e. inflections.

Note that since we are treating everything as symbols with full meaning, it is only sufficient to identify an approximate stem portion that corresponds to lexicon and need not be a strict dictionary entry. Hence it is okay to identify the approximate orthographic portions such as 'enk', 'iruk' as symbols

instead of the strictly dictionary entries such as ‘engu’ or ‘iru’. Knowing the approximate zones of symbols and their functional mapping alone is sufficient for arriving at a full parsed output. Retrieving the actual dictionary entry is a task that is not directly relevant to understand the construction schemas underlying the sentence. For example, as a non-native speaker of Telugu if I come across a token such as *kaDalekkaDundi* ‘Where is the sea?’. I would guess that there are four symbolic units here based on my limited exposure to Telugu namely *kaDal, ekkaDa, un, di*. Strictly speaking, the lexical entry is not *kaDal* in Telugu but *kaDali*. Even without that knowledge, it is perfectly possible for me to guess these above units because I already have a model of *ekkaDa*, *un*, *di* as meaningful symbols and therefore the stem portion *kaDal* should be some other new lexicon symbol. Further based on the sequence of such meaningful symbols, I can directly predict the construction schema labels. For instance if I come across a token ‘ekkaDatagilindi’ and through my limited exposure I have with the language, I guess *ekkaDa* and *tagil* are two lexical symbols, followed by the inflections ‘in, di’. I immediately recognize it must be a *complete schema* in discourse because of the finite inflection. In fact my identification of lexicon is wrong since the actual lexical entry is ‘tagulu’ and not ‘tagil’. However, it is not directly relevant to identify the construction boundaries and their schemas.

It is this intuition that we want to exploit in our splitter and construction labeller modules. The raw text is split into sequence of syllables and given to a *splitter* module. The splitter takes a sequence of syllables and predicts which sequence of syllables should be considered as one meaningful symbol. This can be modelled as a simple sequence labelling task where at any given time the current syllable is grouped as a part of the existing symbol or treated as new symbol. This is shown below: When a syllable

சம்	RT-B
மன்	RT-I
null	DL
அன	RT-B
◌ுப்	RT-I
பி	SF
null	DL
கண்	RT-B
டன	RT-I
அம்	RT-I
null	DL
தெர	RT-B
◌ிவ	RT-I
◌ிக்	RT-I
க	SF
null	DL
முட	RT-B
◌ிவ	RT-I
◌ு	RT-I
null	DL
செய்	RT-B
துள்	RT-B
ளத	SF
◌ு	SF
null	DL

Figure 6.1 Sample splitter output

is considered as beginning a new lexical item, it should be tagged as RT-B by the splitter module; when a syllable is considered as continuation of existing lexical item it is tagged as RT-I and if the current syllable is an inflection of some stem it is tagged as SF. In this way the whole text is split into lexical items and suffixes.

Note that in this approach, we do not worry whether the sandhi that makes up the word is external or internal sandhi because every inflection external or internal is fully functional and meaningful. Thus while conventionally, an expression like 'vantukoNTirukkiRAn' (He is coming) is treated like an internal sandhi of main verb and auxiliary verb and thus should not be sandhi split, our splitter will split it into its lexicon and inflection pattern as follows:

va - RT-B
nt - SF
u - SF
koN - RT-B
Tir - RT-B
uk - RT-I
kiR - SF
An - SF

Thus essentially from the output of splitter module we see that the input token is split into three construction units $va\ nd|u,\ koN,\ Diruk\ kiR|An.$ This output is sent to a construction labeller module which takes a set of lexical stems and their inflections and predicts their construction schema labels directly as follows:

```
va nt|u PR_CONJ
koN UNK PR_CONJ
Tiruk kiR|An PR_FIN
```

Because we have already shown in our theoretical discussions that construction schemas are directly expressed as morphological inflections in Dravidian languages, a CRF sequence labelling of the stem and its inflections along with the contextual window of neighbouring words can predict the above construction schema labels. Note that we do not worry whether 'koN' is an actual lexical item or whether it is a main verb or auxiliary verb and so on. We are at this stage of analysis only bothered about identifying the construction template that orthographic symbols are invoking. After the 'splitter' and 'construction labeller' modules, we need to identify which of these construction templates should be grouped together as one chunk. This is intended to handle what we theoretically had discussed as 'combinative schema' in chapters 2 and 3. Our policy decision was that noun compound MWEs, determiners/ numeric quantifiers modifying a noun alone are the constructions that should be learnt as one chunk in this module. Expressions such as adjectives modifying nouns, quantifiers describing noun or verb, adverbs modifying

verbs, auxiliary verbs following main verbs, serial verb constructions are all treated as separate chunks knowing that we now understand the peculiarity of Dravidian syntax which creates verbiness in all these scenarios and thus allows arguments to be created for even *status* verbs.

Take the input raw sentence 'pakkattu (proximity) vITTu (house) rAman (Ram) vantukoNTirukki-RAn (he is coming)' i.e. 'Near-house-Ram is coming' meaning 'Ram who stays near my home is coming'. After chunker module an output like the following is produced.

```
pakkat tu NN_COMB GC
vIT Tu NN_COMB GC
rAman UNK NN GF
va nt|u PR_CONJ GF
koN UNK PR_CONJ GF
Tiruk kiR|An PR_FIN GF
```

Hence all the units which invoke the 'combinative' schema are identified as GC and the construction unit which forms the head is labelled as GF. In the above example, *pakkattu vITTu rAman* 'Near home Ram' should be identified as one multi-word expression created on the fly to refer to 'a guy named Ram who stays near (my) house' and therefore only 'rAman' becomes GF and the two nouns preceding it are identified as GC. Finally only the construction units which are GFs are sent to the MALT parser to learn dependency relations between these chunks. In this case, we send four chunks to MALT parser namely: *rAman, vantu, koN, TirukkirAn* with all the features we collected from various layers as follows:

```
1 rAman _ NN NN UNK _ _ _ _
2 va _ VB PR_CONJ nt|u _ _ _ _
3 koN _ VB PR_CONJ UNK _ _ _ _
4 Tiruk _ VB PR_FIN kiR|An _ _ _ _
```

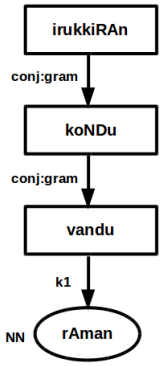


Figure 6.2 Final Full parser Output

The final dependency relation that is learnt for the input is shown in figure 6.2. Note that the main-auxiliary verb sequence is treated as separate chunks and a conjunctive schema relation with the inception state of the main verb is indicated by a label 'conj:gram'. This kind of analysis is useful since these kind of main-auxiliary verb sequences are not merely frozen expressions but are actually dynamic conceptualizations: the sequence can be interspersed by particles in various positions, can exhibit different scopes of negation, echo constructions can be made and so on. One interesting example to show the effect of these phenomena is the sentence: *varAma kirAma iruntu tolaicciTa pORAn* 'Lest he may end up not coming or something of that sort' where negation, reduplication of main verb, conjunctive, infinitive schemas between the interacting processes finally give rise to the above interpretation. Thus the pipeline for the full-parser system that we propose is as shown in figure 6.3. Although through gold

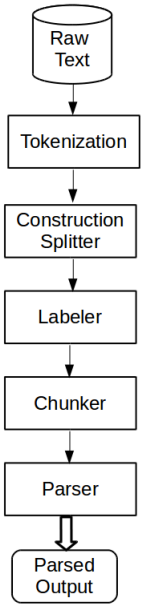


Figure 6.3 Full parser pipeline

annotation experiments, we saw that the labelled attachment score of the annotation scheme is 82.21%, the pipeline accuracy could be much lower due to the pipeline errors, lack of automatic introduction of *null* nodes where required and so on. We implemented a full-parser pipeline with the *construction splitter, construction labeller* and *chunker/ grouper* with small training data size of 2855 words (split as 12000 syllables), 6637 words and 4359 words respectively. For testing, we chose a small set of 60 sentences whose average number of chunks were 7.4. The accuracy of splitter, labeller and chunker modules were 96.81%, 89.97% and 86.56% respectively for this small testing set. While a detailed analysis of every stage of the pipeline can be made with precision, recall and f-measure for every construction schema that would be a larger discussion which is not what the thesis is intending to address. Through this discussion we only want to point out that a fully-working dependency parser could be created and indeed construction schemas could be directly learnt from morphological inflections with

reasonable accuracies. More detailed engineering of the pipeline can be an extension of the current work. We are releasing this full parser as a tool available for download from LTRC website.

Chapter 7

Conclusions

In this work, we have discussed that by exploiting the morphological regularities in Tamil and by mapping these morphological forms to meaning on the theoretical basis of Construction Grammar, we are better able to learn the morpho-syntactic peculiarities in Tamil data. Since these 'construals' are actually learned through morphological features consistently across 5 folds for varying training data size, it shows that there is a merit in applying form-function pairing as a means of syntactic analysis. We have discussed at length all the relevant theoretical perspectives related to the meaningful characterization of Dravidian syntax. One important generalization is that a fully functional characterization of syntax cannot stop with karaka roles alone, but with discourse concepts as well. With gold annotation experiments, we could demonstrate an improvement of 10-12% in LAS consistently for varying data sizes. We have implemented a full parser pipeline that automatically learns construction boundaries, construction schema labels, chunks etc. from the raw text and sends the resulting output of the pipeline to MALT parser for learning dependencies. This idea of form-function pairing can be experimented for other morphologically rich languages as well, since most meaningful information about the syntax comes from morphology in these languages. Another possibility is to explore unsupervised approaches to parsing by making use of these theoretical insights.

Related Publications

Vigneshwaran Muralidaran and Dipti Misra Sharma (2016, April). Construction Grammar based annotation scheme for parsing Tamil. Springer LNCS proceedings of Conference on Intelligent Text Processing and Computational Linguistics 2016.

Vigneshwaran Muralidaran, Ganesh Katrapati and Dipti Misra Sharma (2016, October). Cognitive Construals underlying grammatical aspects and modalities in Dravidian Languages. International Conference on Construction Grammar 2016.

Bibliography

[1] B. R. Ambati, P. Gadde, and K. Jindal. Experiments in indian language dependency parsing. *Proceedings of the ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, pages 32–37, 2009.

[2] B. R. Ambati, S. Husain, J. Nivre, and R. Sangal. On the role of morphosyntactic features in hindi dependency parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 94–102. Association for Computational Linguistics, 2010.

[3] R. Amritavalli. The origins of functional and lexical categories: Tense–aspect and adjectives in dravidian. *Nanzan Linguistics*, 4:1–20, 2008.

[4] R. Amritavalli. Separating tense and finiteness: anchoring in dravidian. *Natural Language & Linguistic Theory*, 32(1):283–306, 2014.

[5] R. Amritavalli and K. Jayaseelan. Finiteness and negation in dravidian. *The Oxford Handbook of Comparative Syntax*, pages 178–220, 2005.

[6] R. Amritavalli and K. A. Jayaseelan. The genesis of syntactic categories and parametric variation. 2003.

[7] M. Andronov. Dravidian languages. *Archiv Orientalni*, 31:177–197, 1963.

[8] E. Annamalai. The variable relation of verbs in sequence in tamil.

[9] A. Bharati, V. Chaitanya, R. Sangal, and K. Ramakrishnamacharyulu. *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi, 1995.

[10] A. Bharati, D. S. S. Husain, L. Bai, R. Begam, and R. Sangal. Anncorra: Treebanks for indian languages, guidelines for annotating hindi treebank (version–2.0), 2009.

[11] A. Bharati and R. Sangal. Parsing free word order languages in the paninian framework. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 105–111. Association for Computational Linguistics, 1993.

[12] A. Bharati, R. Sangal, D. M. Sharma, and L. Bai. Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages. *LTRC-TR31*, 2006.

[13] R. A. Bhat, I. A. Bhat, and D. M. Sharam. Improving dependency parsing of hindi and urdu by modeling syntactically relevant phenomena. *ACM Transactions on Asian and Low-Resource Language Information Processing (under review)*.

[14] R. A. Bhat, R. Bhatt, A. Farudi, P. Klassen, B. Narasimhan, M. Palmer, O. Rambow, D. M. Sharma, A. Vaidya, S. R. Vishnu, et al. *The Hindi/Urdu Treebank Project*. Springer Press.

[15] N. Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[16] W. Croft. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand, 2001.

[17] M. M. Deshpande. Semantics of karakas in pan. ini: An exploration of philosophical and linguistical issues. *Sanskrit and Related Studies: Contemporary Researches and Reflections*, pages 33–57.

[18] R. Dirven. Major strands in cognitive linguistics. *Cognitive Linguistics: Internal dynamics and interdisciplinary interaction*, pages 17–68, 2005.

[19] S. Eggins. *Introduction to systemic functional linguistics*. A&C Black, 2004.

[20] A. E. Goldberg. *Construction grammar*. Wiley Online Library, 2002.

[21] K. Hengeveld, J. L. Mackenzie, et al. Functional discourse grammar. *Encyclopedia of language and linguistics*, 4:668–676, 2008.

[22] S. C. Herring. Aspect as a discourse category in tamil. In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, 2011.

[23] S. Husain. Dependency parsers for indian languages. *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, 2009.

[24] D. Hymes. Prague functionalism. *American Anthropologist*, 84(2):398–399, 1982.

[25] R. Jackendorf. *Patterns in the mind: Language and human nature*. Basic Books, 2008.

[26] K. Jayaseelan. The serial verb construction in malayalam. In *Clause structure in South Asian languages*, pages 67–91. Springer, 2004.

[27] K. Jayaseelan. Coordination, relativization and finiteness in dravidian. *Natural Language & Linguistic Theory*, 32(1):191–211, 2014.

[28] S. Karmakar and R. Kasturirangan. Cognitive processes underlying the meaning of complex predicates and serial verbs from the perspective of individuating and ordering situations in bānlā. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia*, pages 81–87. ACM, 2010.

[29] P. Kiparsky and J. F. Staal. Syntactic and semantic relations in pāini. *Foundations of Language*, pages 83–117, 1969.

[30] S. Kolachina, D. M. Sharma, P. Gadde, M. Vijay, R. Sangal, and A. Bharati. External sandhi and its relevance to syntactic treebanking. *Polibits*, (43):67–74, 2011.

[31] B. Krishnamurti. *The Dravidian Languages*. Cambridge University Press, 2003.

[32] B. V. S. Kumari and R. R. Rao. Hindi dependency parsing using a combined model of malt and mst. In *24th International Conference on Computational Linguistics*, page 171. Citeseer, 2012.

[33] R. W. Langacker. *Language and its structure; some fundamental linguistic concepts: by Ronald W. Langacker*. Harcourt, Brace & World, 1968.

[34] R. W. Langacker. An introduction to cognitive grammar. *Cognitive science*, 10(1):1–40, 1986.

[35] R. W. Langacker. *Cognitive grammar: A basic introduction*. Oxford University Press, 2008.

[36] P. A. Luelsdorff. *The Prague School of structural and functional linguistics*, volume 41. John Benjamins Publishing, 1994.

[37] P. Mannem. Bidirectional dependency parser for hindi, telugu and bangla. *Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, India*, 2009.

[38] T. McFadden and S. Sundaresan. Finiteness in south asian languages: an introduction. *Natural Language & Linguistic Theory*, 32(1):1–27, 2014.

[39] I. A. Melčuk. *Dependency syntax: theory and practice*. SUNY Press, 1988.

[40] M. Menon. Property concepts and the apparent lack of adjectives in dravidian. *The Lexicon Syntax Interface: Perspectives from South Asian languages*, 209:25, 2014.

[41] J. Nichols. Functional theories of grammar. *Annual review of Anthropology*, pages 97–117, 1984.

[42] J. Nivre. Parsing indian languages with maltparser. *In Proceedings of the ICON09 NLP Tools Contest: Indian Language Dependency Parsing*, pages 12–18, 2009.

[43] L. Ramasamy and Z. Žabokrtskỳ. Tamil dependency parsing: results using rule based and corpus based approaches. In *Computational Linguistics and Intelligent Text Processing*, pages 82–95. Springer, 2011.

[44] D. Seddah, R. Tsarfaty, S. Kübler, M. Candito, J. Choi, R. Farkas, J. Foster, I. Goenaga, K. Gojenola, Y. Goldberg, et al. Overview of the spmrl 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. Association for Computational Linguistics, 2013.

[45] S. M. Shieber. *Evidence against the context-freeness of natural language*. Springer, 1987.

[46] S. B. Steever. *The serial verb formation in the Dravidian languages*, volume 4. Motilal Banarsidass Publ., 1988.

[47] S. B. Steever. On the etymology of the present tense in tamil. *Journal of the American Oriental Society*, pages 237–254, 1989.

[48] S. B. Steever. *The Dravidian Languages*. Routledge, 2015.

[49] M. Straka, J. Hajic, J. Straková, and J. Hajic jr. Parsing universal dependency treebanks using neural networks and search-based oracle. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 208, 2014.

[50] A. Szabolcsi. What do quantifier particles do? *Linguistics and Philosophy*, 38(2):159–204, 2015.

[51] R. D. Van Valin Jr. *Advances in role and reference grammar*, volume 82. John Benjamins Publishing, 1992.

[52] K. V. Zvelebil. The present tense morph in tamil. *Journal of the American Oriental Society*, pages 442–445, 1971.