Universal Clause Structure Grammar and the Syntax of Relatively Free Word Order Languages*

K. Narayana Murthy
Department of Computer and Information Sciences
University of Hyderabad
Hyderabad - 500 046, INDIA
email: knmcs@uohyd.ernet.in

Abstract

In this paper we outline our theory of syntactic analysis and apply it to the syntax of Relatively Free Word Order Languages taking the specific case of Kannada. Our theory of syntactic analysis and the associated grammar formalism has been named *Universal Clause Structure Grammar (UCSG)*. The primary objective of UCSG has been to develop a computationally viable grammar formalism that offers insights into the universal features underlying apparently two quite distinct classes of languages - the positional and the relatively free word order (rfwo) languages. While it has been well recognized that constraints on linear position are much weaker in the rfwo languages compared to the so called positional languages, the underlying similarities and real differences between these two classes of languages do not seem to have been understood well. UCSG focuses on different kinds of constraints imposed by grammars, and their relation to computational complexity of parsing. This approach helps us to see vividly the real similarities and differences between the two classes of

*The research work reported in this paper was supported in part by the Department of Science and Technology under grant no. SR/OY/E-10/94

languages. It also leads us to a modular framework for syntactic analysis, making the grammars easier to write and the parsers very efficient. We make no cognitive claims, however.

1 Introduction

This paper is about a theory of syntactic analysis. The job of syntactic analyzer is to accept a sentence in natural language and to produce a description of its internal structure in the light of the given *grammar* - a formal specification of all the valid structures in that language. The grammaticality of the given sentence will also get verified in the process.

There are several aspects of structure that a syntactic theory may be expected to deal with, including assignment of functional roles to the various constituents, analyzing the modifier-modified relationships, resolution of anaphoric and other kinds of references, attachment of prepositional phrases and subordinate clauses, and analysis of emphasis, focus, topic etc. However, it must be made clear that many of these are really issues in semantics or discourse and the contribution of syntax is quite limited. Thus while modifier-modified structures may involve some syntactic constraints, the full problem, namely, determination of exactly what modifies what, and what exactly is the semantic nature of such modification, is well beyond syntax. Similarly, while syntax can provide some constraints for reference, it is far beyond syntax to determine what refers to what and what exactly is the nature of relationship between the reference and the referent. See (Hirst 1981) for example. Syntax can also at most provide some biases or preferences for attachment problems. In UCSG we therefore believe that the first and the most important task for a syntactic analyzer is to assign functional roles to the various constituents in a sentence. This is not only a problem well within the scope of syntax, but also a very basic and essential task for any syntactic system. After all, understanding a sentence includes, at the very least, finding out who did what to whom, where, when, why and so on. We call this aspect of syntactic analysis as functional structure analysis. Hence our primary concern in this paper will be on assigning functional roles to the various constituents in the different clauses within a sentence.

With this perspective in mind, let us now look at the syntax of rfwo languages, taking examples from Kannada. Kannada is one of the four major literary languages of the Dravidian family. See (Murthy 1996b) for application of UCSG for parsing Telugu sentences, another major Dravidian language. We will also compare and contrast with English to bring out the main merits of UCSG. In particular, we will get to see the underlying universals between positional and rfwo languages. We will also get a syntactic analysis framework which leads to simple grammars and very efficient parsing for both positional and rfwo languages.

2 Functional Structure

Consider the following Kannada sentences:

```
K-1a) bha:ratavu inglenDannu  mumbaiyalli so:lisitu
      India-nom  England-acc  Mumbai-loc  defeat-past-3p-n-sl

      'India defeated England in Mumbai'

K-1b) inglenDannu  bha:ratavu mumbaiyalli so:lisitu
      England-acc  India-nom  Mumbai-loc  defeat-past-3p-n-sl

      'India defeated England in Mumbai'

K-1c) bha:ratavu mumbaiyalli inglenDannu  so:lisitu
      India-nom  Mumbai-loc  England-acc  defeat-past-3p-n-sl

      'India defeated England in Mumbai'

K-1d) mumbaiyalli bha:ratavu inglenDannu  so:lisitu
      Mumbai-loc  India-nom  England-acc  defeat-past-3p-n-sl

      'India defeated England in Mumbai'
```

All the sentences above certainly do not mean exactly the same thing. Yet it is clear that they are all equivalent as far as the functional structure of the sentence is concerned - the subject is 'bha:rata', the object is 'inglenD' and the locative is 'mumbai' in all the four cases. In relatively free word order languages, of which Indian languages are instances, word order does not determine functional structure. However, it is not enough to note that these languages permit scrambling. What is important is the fact that in rfwo languages, functional structure information is carried primarily in the inflections and the pre/post positions. Word order is reserved for secondary purposes such as emphasis, focus or topicalization. On the contrary, in positional languages such as English, linear position of constituents plays a major role in determining functional structure. Changing the order most often leads to significant changes in meaning or even renders the sentence ungrammatical. The following English versions of the above Kannada sentences show that even non-arguments cannot be moved about as freely as is possible in rfwo languages.

E-1a) India defeated England in Mumbai
E-1b) England defeated India in Mumbai (!)
E-1c) ?India in Mumbai defeated England
E-1d) ?India defeated in Mumbai England

Word order plays such an all important role in positional languages like English that linguists whose works are primarily based on data from such languages naturally tend to embed linear position at a very fundamental level in their theories. Linear position is such an integral and inseparable part of these theories that such theories can only deal with functional structure in rfwo languages in a highly unnatural, indirect and extremely inefficient manner.

Grammatical functional roles such as subject and object are often defined in terms of dominance and precedence. Thus subject of a sentence is often defined as the noun phrase occurring to the left of the verb and immediately dominated by s. That such definitions are not appropriate becomes clear when we look at data from rfwo languages. Kannada, for example, is a verb final language and all the noun phrases in the sentence normally appear to the left of the verb. The subject noun phrase may

also appear in many different positions relative to other noun phrases in the sentence as we have already seen in sentences K-1a) to K-1d) above. Even the appropriateness of the notion of a vp for rfwo languages can be questioned and hence dominance is also not a sure test. In fact we believe that it is inappropriate to base any of our definitions directly on phrase structure rules or on the corresponding ordered trees. We believe that functional roles should be defined from a more semantic point of view but the definitions should be such that the various functional roles in a sentence can be determined primarily from syntactic constraints only. See (Murthy 1995) for more on functional roles and their definitions.

Of course, it is true that there is a normal or unmarked word order even in the so called rfwo languages. Since linguistic theories are normally presented from an abstract generative point of view, linguists naturally argue that sentences can initially be generated in their default unmarked order and movement rules can then be applied to incorporate topicalization, emphasis or such other things. However, the real life problem faced by Natural Language Processing is often that of analyzing and interpreting natural language sentences as they are used by people. From an analysis point of view, it would not at all be appropriate to first analyze topicalization etc. and get the sentence into its unmarked word order and then analyze the sentence for its functional structure. Word order is not significant for functional structure analysis and the simplest, the most straight forward and the best approach is to simply not to consider word order at all for this task.

If linear position does not hold the requisite information for functional structure analysis we must ask the question where else is the required information? For rfwo languages the clue seems to be in the inflections and post positional markers. Thus morphology plays an extremely important role in the analysis of rfwo languages. See (Sridhar 1990) for more on morphology of the Kannada language. Let us now look at the basics of Kannada syntax. A detailed study of all aspects of Kannada sentence structure is beyond the scope of this paper. (Sridhar 1990) is a very good introduction to all aspects of Kannada grammar. It also gives a large number of references to other original works of interest.

Kannada is a S-O-V language. That is, the default or unmarked order of constituents is Subject first, then the Object and finally the verb. However, Kannada,

being a relatively free word order language, permits substantial amount of freedom in the order of constituents although normally the verb remains in the sentence final position. Word order becomes less important mainly because noun groups are marked for cases and the verb agrees with the subject in gender, number and person. In fact, subjects and objects are often dropped. There are copular sentences where the verb is often not shown overtly. Normally all modifiers precede the modified. There are a variety of subordinate clauses. Subordinate clauses also precede the main clause. They typically involve special non-finite forms of verbs and are marked by a variety of particles called sentinels in UCSG. Sentinels occur invariably in the clause final position and mark the right hand boundary of the respective clauses. All these assertions are to be taken as rules - there are exceptional situations where deviations from these rules are possible. Also, most of these rules apply not only to Kannada but to Dravidian languages in general.

The subject of a sentence is expressed by a noun group in the nominative case. The nominative case marker is empty. Kannada also permits dative subject constructions where the understood subject is indicated by a noun group in dative case whereas the surface subject appears in the nominative case. See K-7) below. Subjects can be dropped. See K-2) and K-5) below.

```
K-2) na:Le     barutte:ne
     tomorrow  come-pres-p1-sl

     'I will come tomorrow'

K-3) magu  ha:lu kuDiyuttide
     child milk  drink-pres-cont-3p-sl-n

     'The child is drinking milk'

K-4) idu  nari
     this fox

     'this is a fox'
```

```
K-5) nannannu baiyabeDa
     me       scold-neg-sl

     'Do not scold me'

K-6) chinnu giDagaLige   ni:ru  ha:kidaLu
     Chinnu plant-pl-dat water  pour-past-p3-f-sl

     'Chinnu watered the plants'

K-7) nanna saikallige mu:ru cakra ide
     my    cycle-dat  three wheel be-p3-n-sl

     'My bicycle has three wheels'
```

The Object is indicated by a noun group in the accusative case. See K-5). However, the case marker can be dropped in most situations. The suffix is mandatory for human nouns and when the direct object noun group carries other suffixes or particles. Dropping of accusative suffix can cause ambiguities since the nominative suffix is also empty. People seem to apply their world knowledge to get the correct interpretation. Heuristics based on certain semantic features such as the animate feature often help. K-3) above is an example of this. When no other information is available, word order can be used as the last resort.

The indirect object is expressed by a noun group marked for the dative case. The dative suffix is also used to indicate other roles including psychological subjects in dative subject constructions as in K-7) above.

There are also other case markers such as the locative and the instrumental/ablative. Also, a variety of post positions are used in combination with inflection to indicate different functional roles. In UCSG we bundle up all sources of functional structure information including morphological inflections, prepositions and post positions as well as linear position and call them 'surface case markers'. All these are simply

different manifestations of one and the same phenomenon. It may be noted that Paninians take a similar view and try to provide an extended notion of 'vibhakti' (Bharati et al 1996) in their Paninian Grammar (PG) framework.

In UCSG, like in PG, we view functional structure analysis as essentially constraint satisfaction - any assignment of functional roles which satisfies all the constraints of the grammar simultaneously, constitutes a valid functional structure. We bring to bear constraints of subcategorization frames and selectional restrictions in a top-down fashion simultaneously with surface case marking information on the potential groups in a bottom-up fashion to obtain the complete functional structure. As far as multi-clause sentences are concerned, however, we show that the UCSG approach is far better than other western grammar formalisms as well as PG, both for positional and rfwo languages.

3 Hierarchical Structure

The main strengths of UCSG lie in the way complex sentences are analyzed. Clauses in a sentence can be nested one inside the other, resulting in a hierarchical or tree-like structure. We call this aspect of structure as hierarchical structure. In UCSG we view a clause as a unit of linguistic structure that corresponds to any one action or a state description. Every clause has exactly one verb group which indicates the action or state being described. Every clause also has its own set of functional roles to describe the action or state in full. *Every clause has its own functional structure.* This is obviously true of both positional and rfwo languages.

Clauses in a sentence are not, however, completely independent of one another - there can be inter-clause dependencies. For example, a noun phrase being modified by a relative clause has two roles to play, one in the relative clause and one in the outer clause. In UCSG we have analyzed inter-clause dependencies in detail and we have shown that *all inter-clause dependencies systematically flow down the clause structure tree from the root towards the leaves* (Murthy 1995).

Also, it has been observed that constituents of a clause do not cross clause boundaries in scrambling. See, (Tirumalesh 1979) on this for Kannada. In UCSG we go much further and make a stronger claim about all languages: *it is an extremely im-*

portant principle of syntax that clauses exhibit well defined clause boundaries and the participants of a clause do not normally cross clause boundaries. Violations of this principle should only be viewed as exceptions to this general rule rather than as evidence for the invalidity of this principle.

Hence if we can determine the clause structure first, including the clause boundaries, we can start our functional structure analysis from the matrix clause and recursively analyze the embedded clauses taking care of all inter-clause dependencies, thereby making functional structure analysis of a clause completely independent of any other clause. As a byproduct, this approach disparages the notion of long distance dependencies. We have shown that all these dependencies are related to hierarchical and functional aspects of structure, they have nothing to do whatever with distance, long or short (Murthy 1995, Murthy 1996a). The only prerequisite would be determining the clause boundaries before and without applying functional structure constraints.

In UCSG we show that it is possible to determine clause boundaries in a very simple and efficient way even before we apply any of the functional structure constraints. Knowing the clause boundaries, we can analyze the functional structure one clause at a time, starting from the matrix clause and recursively analyzing the embedded clauses, passing down information and expectations about inter-clause dependencies as we move down the clause structure hierarchy. This strategy of *working from whole to part* makes functional structure analysis of a clause completely independent of any other clause. *Functional structure is essentially local to a clause.* It may be noted that none of the theories of syntactic analysis available so far have given us a method of exploiting this locality of functional structure. All the available parsing models analyze the entire sentence as a unit. This makes parsing computationally very complex - a lot more complex than need be. Our approach leads to very efficient parsing for both positional and rfwo languages. We also get significant insights into the underlying universals between positional and rfwo languages.

We observe that the crucial pieces of information required for analyzing clause structure lie in verb groups and certain functional words or markers called sentinels. We show that verb groups and sentinels contain all the required information for recognizing clauses, for determining the nested or hierarchical structure of clauses as

well as for determining the clause boundaries, although only partially.

We observe that every clause in a sentence except for the main clause has a sentinel marking one of the boundaries of that clause. The sentinel parametrically marks either the beginning or the end of the clause depending upon the language in use. Also, by definition, every clause must have exactly one verb group. Thus verb groups and sentinels behave like brackets and impose very strong constraints - the brackets must match properly. Thus the total number of verb groups in a sentence must be exactly one more than the total number of sentinels. If we start with a count of 1 and scan a Kannada sentence from left to right, incrementing the count whenever we get a sentinel and decrementing it on encountering a verb group, the right most location where the count becomes zero indicates the matrix clause verb group. The final value of the count must be zero for the sentence to be well formed. *Constraints on clause structure imposed by verb groups and sentinels are thus very strong yet very easy to apply.* In fact, these constraints also help us in reducing lexical ambiguities to some extent, especially the more critical ambiguities such as noun/verb and sentinel/non-sentinel ambiguities.

On the other hand, other units of a sentence such as noun phrases provide only secondary and weak constraints. Weak constraints are less useful but require more effort to apply. It is better to identify and apply only a few very strong constraints to begin with than to simply apply all the constraints of grammar without regard to their strength. In UCSG we analyze the structure of clauses in a sentence purely based on verb groups and sentinels, to the total exclusion of all other constituents. Consider the following sentence:

```
K-8)  ba:nkannu do:cabe:kendukonDidda          Daka:yitaru
      bank-acc  rob-inf-want-quot-refl-rel-past dacoit-pl

      ra:tri aDaviyoLakke      para:riya:girabe:kendu
      night  forest-inside-dat escape-completive-be-inf-want-quot

      po:li:saru nambidaru
      police-pl  believe-past-3p-pl
```

```
'The police believed that the dacoits who wanted to rob
 the bank must have escaped into the forest at night'
```

In the above Kannada sentence, the relative marker 'idda' indicates the end of the relative clause and the so called quotative 'endu' marks the end of the clausal subject of the main clause. Correspondingly, the English sentence in the gloss above has the relative pronoun 'who' marking the beginning of the relative clause and the complementizer 'that' indicating the start of the clausal subject of the main clause. We believe that the most important function of such words or markers is in the exposition of the clause structure of the given sentence.

In UCSG we distinguish between clausal subjects, clausal objects and adverbial clauses etc. on the one hand, and relative clauses on the other. The former class of clauses are role fillers, they take on a functional role on their own in an outer clause. These clauses are denoted by the symbol 'sbcls'. Relative clauses, on the other hand, modify some constituent expressed by say, a noun phrase, which is really the basic role filler. We denote relative clauses by the symbol 'rlcls'. Note that for the purpose of hierarchical structure analysis, we do not have to worry about the internal structure of clauses.

Kannada has a variety of clauses. Kannada is a verb final language and both rlcls and sbcls can appear only to the left of the main clause verb group. Also, the sentinels denoted 'rl' and 'sb' respectively mark the end points of rlcls and sbcls in Kannada. Keeping these things in mind, we give below a set of phrase structure rules for clause structure analysis in Kannada. Applying these rules on the verb group - sentinel sequence, we obtain the clause structure tree as depicted in Figure 1.

```
do:cabe:kenduko:+idda para:riya:girabe:ku+endu nambidaru
vg               rl   vg                  sb   vg


s      ==>   (sbcls)*  fcls
fcls   ==>   (rlcls)*  vg
rlcls  ==>   s         rl
sbcls  ==>   s         sb
```

In this figure, each node in the tree is annotated with three numbers which relate to the boundaries of the corresponding constituents. The n^{th} word in the sentence is supposed to lie between positions n and n+1. For the purpose of indicating positions in the sentence, morphemic sentinels are treated like words. Of primary interest to us are the rlcls and the sbcls. The left most number indicates the earliest position in the sentence where the clause can begin and the middle number indicates the right most position in the sentence where the clause can begin. The third number indicates the exact position where the clause ends. Thus in this sentence the rlcls begins somewhere between positions 1 and 2 and terminates at 4. The sbcls starts somewhere between 1 and 2 and ends at 9. This is also indicated by the bracketed structure in which we have used curly braces to indicate sbcls and square brackets to indicate rlcls.

The part of the sentence lying between the first number and the third indicates the *local domain* of the clause, beyond which we are sure that the clause does not stretch. To analyze the functional structure of any clause we need not even consider any constituent lying outside its local domain. We also need not consider constituents lying entirely within the local domains of clauses which are embedded within the local domain of the clause in question since clauses can take roles in other clauses only as a whole.

The part of the sentence lying between the first and the second numbers is called the *grey area* for that clause. We cannot say for sure whether constituents in the grey area really belong to the clause in question or not. This is the price we have to pay for having ignored all constituents but for the verb groups and sentinels for clause structure analysis. In the example sentence, only one word - 'ba:nkannu' lies in the grey area. Until we apply functional structure constraints, we cannot be sure whether this word is part of the 'nambidaru' clause or 'do:cabe:kenduko:' clause or the 'para:riya:girabe:ku' clause. We are very sure, however, that 'Daka:yitaru', 'ra:tri' and 'aDaviyoLakke' can only be part of the 'para:riya:girabe:ku' clause and 'po:li:saru' can only be part of 'nambidaru' clause. There is thus a substantial advantage in terms of localization of functional structure analysis. While the sentence itself has 10 words (including morphemic sentinels), the parser had to look at only the verb group - sentinel sequence of length 5. There are also only 4 rules in the clause structure grammar. In fact the rules are all of the Context Free Grammar (CFG)

power (Murthy 1995, Murthy 1996a). Hence determining clause structure is itself a very efficient process. Hierarchical structure analysis in UCSG has a worst case time complexity of $O(n'^3)$ where n' is the length of the verb group - sentinel sequence, which is typically much smaller than the length of the original sentence.

However, the most significant aspect of this approach is that the *grammar of clause structure is universal* - parametric variations of the same grammar rules apply for positional languages like English as well. In English the sentinels mark the starting positions of clauses. Also, English being a 'S V O' language, rlcls may be found on either side of the verb group of a clause. The sbcls may also appear either to the left or to the right of a clause. Hence the set of rules for analyzing the clause structure of English sentences would be

```
s      ==>   (sbcls)*  fcls (sbcls)*
fcls   ==>   (rlcls)*  vg   (rlcls)*
rlcls  ==>   rl         s
sbcls  ==>   sb         s
```

No other changes are required. With these four rules and following exactly the same procedure as for Kannada we can obtain the following clause structure for the English sentence in the gloss of K-8) above. The only parametric difference is that in English the grey areas lie towards the right end. For each clause, we would get a single definite starting position but two limiting positions for the far end of the clause.

```
The police believed { that the dacoits [ who wanted ] to rob
the bank ] must have escaped } into the forest at night }
```

Thus we find that *clause structure is the most significant underlying universal feature. Freedom of word order is mainly a clause internal phenomenon.* As far as

hierarchical structure of clauses is concerned, positional and rfwo languages are identical. Hence the name *Universal Clause Structure Grammar (UCSG)*. Note that we could make this crucial discovery about what is really common between positional and rfwo languages only because analysis of clause structure was carried out solely based on verb groups and sentinels to the total exclusion of all other elements, especially the noun groups. This substantiates our hypothesis that the primary carriers of clause structure information are the verb groups and the sentinels. Freedom of word order is mainly to do with noun groups and such other elements. Other theories of syntax have all failed to discern this universal aspect of language mainly because they fail to make a clear distinction between the strong, primary constraints and the secondary and weak constraints. In all these theories of syntactic analysis, a sentence has to be analyzed as a whole and this makes both the grammar and the parser much more complex than really necessary. *UCSG is unique in suggesting an independent hierarchical structure analysis, carried out before and without applying any of the functional structure constraints. Localization of functional structure analysis is thus possible only in UCSG.*

To be fair, we have to admit that there are important exceptions to what all we have seen above. There are some kinds of clauses where no overt sentinels are used. Infinitival and gerundial clauses in English are examples. One possible approach would be to propose additional (possibly empty) sentinels and include these kinds of clauses also for hierarchical structure analysis. Alternatively, one may simply exclude these clauses from the hierarchical analysis stage, deferring their analysis to the functional structure analysis phase. We observe that infinitival and gerundial clauses are nominalized clauses. They exhibit dual nature - they share properties of clauses as well as of noun groups. Hence they can very well be treated like noun groups as far as hierarchical structure analysis is concerned. What would be lost is some ability to localize functional structure analysis further and what would be gained would be simplicity and efficiency of hierarchical structure analysis.

There are also structures, which we call *reduced constructions*, where sentinels and/or part of verb groups may be optionally dropped out. These reduced constructions mean exactly the same as their complete counterparts and hence UCSG takes the view that reduced constructions should be handled by the same rules of grammar

as their full counterparts. Introducing additional rules in the grammar to take care of these special cases is neither essential nor desirable. However, in order to use the same grammar rules as for normal constructions, the syntactic analyzer must be able to recognize such reduced constructions, identify what is really dropped out and where. After inserting the optionally dropped out items, parsing can proceed as usual. After all, distinguishing between the rule and the exception, identifying exceptional situations and realizing how exactly a particular exceptional situation differs from the normal or the standard one, are all basic features of human intelligence. See (Murthy 1995) for more on this.

4 Linear Structure

In UCSG we observe that the fillers of functional roles are either entire clauses or groups of words. In UCSG we analyze the clause structure first and take full advantage of locality of functional structure. Let us now see how we can identify groups of words that can take up various functional roles in a clause. These *groups of words behave as atomic units at both hierarchical and functional structure analyses phases and hence it is best to recognize these groups right in the beginning.* The rest of syntactic analysis can then deal only with whole groups of words rather than with individual words. *The effective length of sentences would be significantly reduced and the analysis becomes more efficient.* It must be noted that other syntactic analysis models deal directly with words, making grammars as well as parsers more complex than need be.

In UCSG we define a *word group* or simply a *group* as a typically contiguous group of words and/or morphemes that can potentially take on one functional role in some sentence. Noun groups, verb groups, adjective groups and adverb groups are some of the important types of groups. Note that in UCSG prepositional groups are equated with noun groups. Some examples of word groups in English and Kannada are given below:

English	Kannada
one small poem	ondu saNNa padya
on the mountain	beTTada me:le
will keep on singing	ha:Duttale: irutta:re

In UCSG we make a clear distinction between word groups (which should ideally be called phrases) and clauses. *Clauses are inherently more complex than word groups.* Clauses may be nested recursively one inside the other several levels deep while by definition word groups do not involve any kind of hierarchical structure. While a noun phrase may include relative clauses, a noun group cannot. We reserve the term verb group to sequences of auxiliary and main verbs only. In contrast verb phrases may even include one or more noun phrases in it.

By definition, there is no hierarchical structure of any kind within word groups. There are only three relevant aspects of structure within word groups. Firstly, some items may be optional. Thus there are noun groups in English with or without a determiner. Some items may be repeatedly used. For example, there may be any number of adjectival modifiers in a noun group. Thirdly, and most importantly, linear order of words is almost invariably significant. It should be emphasized that *linear order of words within word groups is equally important in positional languages like English and rfwo languages.* See (Tirumalesh 1979) for restrictions on scrambling in Kannada. Also, we have shown that *all potential word groups in a sentence can be obtained in a single linear scan of the sentence and in linear time using Finite State Machine power. Hence the name linear structure.* See (Murthy 1995) for a more detailed discussion on the various issues involved in linear structure analysis.

5 Conclusions

In the past, when grammars of a particular complexity class were believed or shown to be inadequate for handling all of the syntactic phenomena, this was often simply taken to imply the need for grammars of the more complex class. We instead asked ourselves which aspects of syntax really require which kinds of grammars. We have thus achieved our dual major objectives of universality and efficiency by diving the task of syntactic analysis into three independent modules - the linear, the hierarchical and the functional modules. This modularity also makes the grammars easier to write. The grammar of linear structure is very simple to write the grammar of hierarchical structure is largely common across many languages. Thus adapting grammars and parsers of one language to another would also be greatly facilitated in UCSG. Also,

as a byproduct of our work-from-whole-to-part strategy, we have gotten rid of all problems of long distance dependencies in an extremely simple and elegant way.

We have for the first time a grammar formalism equally well suited for positional and rfwo languages. UCSG is also much more efficient than other grammar formalisms. in UCSG word groups are first identified in linear time, the fastest possible, and rest of the analysis is done using these groups of words as atomic units. This by itself makes parsing a lot more efficient than if we directly dealing with words. Also, we are able to localize the functional structure analysis to the respective local domains of the clauses, thereby making functional structure analysis very efficient. Hierarchical structure of clauses, which enables us to take advantage of locality of functional structure, is itself carried out in cubic time using a very small number of CFG rules on only the verb groups and sentinels in the sentence. *Localization of functional structure analysis makes UCSG superior to and more efficient than Paninian Grammar, even for rfwo languages.*

UCSG parsers have been implemented for English, Telugu and Kannada. Work is currently on to enhance the coverage and robustness of these parsers. UCSG systems are currently being applied for spell checking and English to Kannada machine assisted translation. UCSG parsers have been successfully applied to the problem of metaphor interpretation also (Varma 1996).

References

Bharati et al 1996

Akshar Bharati, Medhavi Bhatia, Vineet Chaitanya, Rajeev Sangal, "Paninian Grammar Framework Applied to English", Tech. Rep TRCS-96-238, CSE, IIT(K), 1996

Hirst 1981

Graeme Hirst, "Anaphora in natural language understanding: A survey" (Lecture Notes in Computer Science 119), Springer Verlag, 1981

Murthy 1995

K. Narayana Murthy, "Universal Clause Structure Grammar", PhD thesis, Dept. of CIS, University of Hyderabad, 1995

Murthy 1996a

K. Narayana Murthy, "Universal Clause Structure Grammar", To appear in the special issue on NLP and ML of Computer Science and Informatics, J. of CSI

Murthy 1996b

K. Narayana Murthy, "Parsing Telugu in the UCSG Formalism", Proc. of the Indian Congress on Knowledge and Language, Vol 2, pp 1-16, Central Institute of Indian Languages, Mysore, January 1996

Sridhar 1990

S. N. Sridhar, "Kannada", Descriptive Grammars series, Routledge, 1990

Tirumalesh 1979

Tirumalesh K V, "Reordering Rules in Kannada and English", Unpublished PhD thesis, Central Institute of English and Foreign Languages, Hyderabad

Varma 1996

Vasudev Varma, A. Sivasankara Reddy, "Knowledge based metaphor interpretation", Knowledge-Based Systems, vol 9, pp 339-342, 1996