

Volume 1:2 April 2001

## LINGUISTICS AND DOCUMENT CLASSIFICATION: A PARADIGM-MERGER APPROACH

**B. A. Sharada, Ph.D.**

---

### 1. INTRODUCTION

The association of linguistics and document classification has given birth to *Infolinguistics*, wherein the document titles are syntactically represented and semantically interpreted (Sharada 1995). Noam Chomsky is considered the architect of the dominant approach in linguistics that is used to explain language. When it comes to an understanding of the index language it is S. R. Ranganathan who is considered to be the architect of the unique analytico-synthetic scheme of classification, namely, Colon Classification (CC) system. The Chomskyan frame is for natural language and the latter is for artificial language, called index language.

### 2. CHOMSKYAN SCHOOL OF THOUGHT

The Chomskyan paradigm has developed since 1950s and has faced many crises. But it has successfully resolved most of these crises and accounts for more important linguistic phenomena. Despite several reservations, the Chomskyan framework has been rather universally accepted.

In the *Aspects* model, Chomsky introduced deep structure, which gave semantic interpretation, and a surface structure, which gave phonetic interpretation of the sentence. The deep structure was later called D- Structure and the surface structure as PF (Phonetic Form). The *Aspects* model was considered a standard theory. 1980s saw Chomsky's famous government and binding theory, which had many rules and sub theories. This was followed by the principles and parameters theory. This was introduced in search of a more elaborate and precise apparatus needed for semantic interpretation. This theory reduced the number of rules and brought the principles of Universal Grammar (UG) and language faculty to the center stage of linguistic research. This minimalist program helped in finding a way towards achieving Universal Grammar. Chomsky puts UG as follows. It is not a grammar by itself. The theory of a particular language is its grammar; "the theory of language and the expressions they generate is Universal Grammar"(Chomsky 1995). He also states that UG is the name for a theory of the initial state of the language faculty.

### 3. PARALLEL THOUGHTS OF THE TWO PARADIGMS

The two components of language are: a lexicon and a computational system. The lexicon specifies the minimal items that enter into the computational system, with their idiosyncratic properties. The computational system uses these elements to generate derivations and structural descriptions. The indexing language also has got two basic elements: Vocabulary and syntax. Vocabulary is a list of terms used in the system and syntax is the recognized pattern of relationship between the terms used in the system. Function of an indexing language is to provide point of access to the seekers of information in addition to communicate the semantic content of its expression in a simple manner.

In the last four decades, we see a close parallel between the developments of the grammar of the indexing language and the studies of the theory of syntax of language. In the indexing language field, S. R. Ranganathan's analytico- synthetic paradigm was developed two decades earlier than the development of Chomskyan theory of syntax. To demonstrate the similarity in thought between these paradigms, a few examples may be mentioned here. Ranganathan introduced *Absolute* syntax for indexing language at the cognition level. Parallel to it, Chomsky stated that grammar is inbuilt at the cognition level, which coincides with the definition of absolute syntax. Ranganathan's analytico synthetic scheme is based on three planes of work, namely, Idea, Verbal, and Notational planes. Chomsky worked in three levels of grammatical analysis in his first generation grammar, namely, Phrase structure, Transformations and morphophonemic. Chomsky defined a sentence as the combination of Noun Phrase and Verb Phrase.  $S = NP + VP$  where  $NP = \text{Determiner} + \text{Noun}$  and  $VP = \text{Verb} + NP$ . In other words,  $S = NP + V + NP$ .

Ranganathan, in his Colon Classification, introduced five fundamental categories and defined their structure as:  $S = [BS][P][M][E][S][T]$  where BS is the Basic subject and the rest PMEST are the fundamental categories. Their full forms are P=Personality, M=Matter, E=Energy, S=Space and T=Time. Here the Energy facet is placed in the middle of the facet syntax. It is seen clearly that Chomsky also placed verb or the action part of a sentence in the middle. It is agreed that there is no verb in indexing language. But noun variants of a verb appear in most of the titles, which has been expressed as Energy facet by Ranganathan.

Ranganathan went on improving his colon classification based on the experience and brought up to seven editions of the scheme introducing depth schedules in different subjects based on the growth in Universe of knowledge. Chomsky also went on improving his models. In search of a more elaborate and precise apparatus needed for semantic interpretation, he introduced many theories. The government and binding theory introduced by him in 1981 was more explicit and more explanatory than the earlier theories.

Regarding Ranganathan's models, based on the subsystems of the principles, a few sub theories or theoretical modules were introduced. These theories were tested on indexing language on the basis of Ranganathan's fundamental categories.

Ranganathan's Theta theory and Case theory are very much suitable to indexing language. All these factors have been discussed in my earlier paper (Sharada 1995).

#### **4. THE MINIMALIST PROGRAM: MARCH TOWARDS UG**

In the 1990s, Chomsky reduced the number of rules in his Principles and Parameters theory. The parameters were set up on the basis of empirical facts about languages to decide whether a particular principle is applicable to a certain type of language or not. According to him,

1. Human beings are the only species to have language. It is embodied or inbuilt as the language faculty in human brain.
2. Language faculty interacts with other faculty in the brain to generate phonetics and logical representation.
3. LF representation (semantic and pragmatic component) is interpretable and visible to the mental eye.
4. Phonology is different in different languages but semantics is same for all languages.
5. Grammar is inbuilt at the cognition level.
6. The language is embedded in performance systems that enable its products to be used for articulating, interpreting, referring, inquiring, reflecting and other actions. Structural design is a set of instructions for these performance systems.

Universal Grammar is natural grammar. Language is the universal property. Language faculty is in the human brain. Language faculty interacts with other faculty in the human brain. Through this both phonetic form and the semantic and pragmatic forms are integrated.

In the case of Index language, the variety of language itself is artificial. In order to bring the concept of "Universal" to index language, we have to identify what the similarities between the index language and the natural language.

Keeping the above points in view, we shall identify the features of the Minimalist program. Minimalist program is not a principle but a way of doing things. It is a program of minimal theory, a search for simplicity, and it tries to reduce the language specific rules including the principles and parameters. It has not only abandoned the traditional concepts of "rules of grammar" and "grammatical construction" but also paved the way for asking questions that has no real counterpart in the earlier theory. The language-specific rules are reduced to these choices. The notion of grammatical construction is eliminated, and with it, construction-specific rules. Constructions such as verb phrase, relative clause, passive, etc., are taken to be taxonomic artifacts, collections of phenomena explained through the interaction of the principles of UG, with the values of parameters fixed. It has no D-structure or S-structure but only LF and PF interface. The LF interaction takes place after lexical items are chosen from the lexicon and the computational system starts building representations. Each lexical item has three features,

namely, Semantics, Phonological, and Syntactic. The generative process consists of merge and move. While doing so, it follows a step-by-step process.

Example: *The book was read.*

1. Step1 - Merge [the] and [book] yielding [the book]
2. Step2 - Move [read] and [the book] yielding [read the book]
3. Step3 - Merge [be] with [read the book] yielding [be read the book] Move [the book] yielding [the book was read t]

Merge AGR [the book was read] Matching of case features, AGR features, tense features, etc. are checked under strict locality conditions. Another recurrent theme has been the role of "Principles of economy" in determining the computations and the structural descriptions (SD) they generate. They are fundamental to the design of language.

Important concept or the key elements in minimalist program is that of merge and move. If merge is not a part of move, it is pure merge. Movement is guided by economy conditions, Which involve economy of derivation and representation. It always takes the shortest route. At each step of derivation the principle of economy allows only a minimum of transformational activity. The uninterpretable features have to be eliminated before semantic representation. Movement takes place as late in the derivation process as possible. It is triggered by the need to license inflected elements. If steps are taken correctly, merged elements converge, if not crash. If they crash, we have ungrammatical structure.

## **5. APPLICATION OF MINIMALIST PROGRAM TO INDEX LANGUAGE**

In order to accommodate future developments in the universe of knowledge, Ranganathan's analytic- synthetic paradigm is recommended. The matching phenomena of Ranganathan's paradigm to that of Chomskyan's have enabled us to apply the minimalist theory to index language (IL). The phonemes of IL are the ordinal numbers in the notational plane, parts of speech are the five fundamental categories and the lexicon are taxonomic or thesaurus-based. The grammar of IL is in the form of postulates and principles by which the ordinal numbers of the descriptors are combined in order to translate the specific subject to class representations. The approach is to proceed towards universal index language (UIL).

## **6. WAYS TO ACHIEVE UIL**

As defined in the minimalist theory, reduce the number of postulates and principles. To achieve the above, simplify the rules. The Minimalist theory in IL should start at the cognitive level. To achieve the above, adopt Ranganathan's absolute syntax. Absolute syntax is defined as the sequence of the component ideas in a subject helpful and acceptable to majority of users. The structure denotes logical form. The parts, properties or aspects are in some manner related to each other (Neelameghan 1971)

## 7. WAYS TO APPLY MINIMALIST THEORY

While analyzing the document title,

1. At the cognition level itself identify the Basic subject and tag the concepts with fundamental categories.

Example: Sociology of middle class alcoholism in the developing countries in 1990s.

The analysis at the absolute syntax level would be as follows:

[Sociology(BS)[of[middle class(P1)[alcoholism(P2)in the[developing countries(S)in[1990s(T)]]]]]]

2. Merge and move - arrange the FCs according to facet syntax in the notational plane.

3. The result of the above interaction and by incorporating any additions such as few devices, the required class number is got. That is, by the application of the minimalist program, the whole traditional system of classification in nine steps are reduced to that of just three steps. This reduction of steps is justified for the reason that from generation to generation considerable improvement is noticed at the cognition level of human beings. Here we may recall Ranganathan's statement that a veteran may not take more than a minute or two in classifying books.

## 8. WAY TO REDUCE THE RULES PART OF IL IN CC

Keeping in view the present-day needs and future avenue, the features of a pre-coordinate indexing system (PCIL) followed in CC is quite ideal. Following are the few recommendations which could reduce and simplify the rules part of CC.

1. The fundamental categories must be well defined and universally acceptable. [The definitions given to PMEST are clear in the schedule. But some time [S] and [T] also act as [P]. Complications are noticed while defining Space and Time facets. For example, the facet formula for journal is ACI[1P1],[1P2] m44,N. In BS history also, History of India, the term *India* does not denote Space isolate but is deemed to be a manifestation of the Personality isolate. This, Ranganathan puts it as pitfalls. It is done to the impersonation of one fundamental category as another. This will create more ambiguities. The fact here is that it should be easily understood by the artificial intelligence. Uniformity has to be maintained throughout the schedule.]
2. Since CC is of analytico synthetic scheme one concept can have only one entry.
3. If only one facet syntax is followed for all subjects one can avoid going through the rules part in each subject.
4. The special libraries and few research libraries can come forward with the developments in their disciplines and if any lacuna found in the CC, which could be corrected (Sharada 1997).

5. The PCIL should be accompanied by relative index wherein each concept may be specified with their placing.
6. The schedule should not be biased, as some of the famous PCILs are criticized as western biased, CC is criticized as eastern biased (Example - Hindu sacred books).
7. The whole schedule has to get updated with latest developments in each subject and provide place for future development.
8. CC has mixed notation and has seventy four symbols used in the notational plane.

The notation of CC has been criticized on the grounds of its length and complexity of class numbers. Specificity and synthesis have combined to give such lengthy class numbers. No doubt, for simple subjects CC gives shorter numbers than other PCILs. The class numbers are confusing only because of mixed notation. Only the indexer can understand the notations and is not user friendly. Hence notation has to be accepted at the universal level. An attempt has to be made to improve the notation and make it user friendly.

For example: Sociolinguistic study of modern Marathi and medieval Hindi speakers in Karnataka in 1981. P(Y),155=xJ&m152=xE.4413'N81. Even when all the above procedures are adopted one need not fear about the volume of the schedule, as it will be done on the digital platform, wherein the content can be more and the access will be handy and user- friendly.

## **9. APPLICATION OF NATURAL LANGUAGE PROCESSING (NLP)**

In the context of the present study, the use of Natural Language Processing (NLP) could be utilized in two ways - The compilation of the schedule, and information retrieval. While developing the parsers in the NLP environment for index language also, two methods can be followed -

- a. The usage of NLP parsers developed using the grammatical categories
- b. Developing IL parsers adopting fundamental categories in place of grammatical categories.

## **10. LP WITH GRAMMATICAL CATEGORIES**

As the index language (IL) does not possess all the qualities of natural language, it has to be modified. The first thing the parser identifies in a sentence is verb phrase. It has already been discussed that instead of verb, a nominal form related to the verb occurs in IL. The requirements and the rules to develop parsers for IL have been discussed in my earlier study in five steps (Sharada 1998). Now the computational system may be designed in such a way as to take representations in natural language and concepts in indexing language and modify them. Then the UG must provide some device to present an array of items from the lexicon/corpora in a form accessible to the computational system. We may take this form to be some version of X-bar theory. That is, any theory, which is suitable to indexing language, could be chosen. The X-bar theory has been

selected because, the concepts of X-bar theory are fundamental and in a minimalist theory the crucial properties and relations will be stated in x-bar theoretic terms. It is a categorical component of grammar and is the phrase structure rule. Its structure is composed of projections of heads selected from the lexicon. In simple terms, 'what can dominate what' is the question. In case of IL, hierarchy will be maintained under each class. Also it can accommodate any number of rounds and levels. The (BS) PMEST structure has revealed a modulated formulation of decreasing concreteness. In the similar way, this theory also reflects the normal mode of communication of ideas and confirm to the sequence of ideas preferred by a majority of specialists in the subject.

Now for NLP, several software programs are readily available and recently neural networks are also adopted. Hence the unification of grammars are based on mathematical, logical and statistical probability models. However it is suggested that, while adopting any structural rules it is better to consult the expert in that field.

## 11. NLP USING FUNDAMENTAL CATEGORIES

- I. Creating the corpora Recently Corpora are any body of the data. In the present context it can be a complete schedule. It should be comprehensive.
- II. Analysis

For the analysis we have to develop parsers. The only difference here is, instead of using the grammatical categories we can use fundamental categories since the language is IL. Since they are abstract parameters, translate them to concrete parameters. Following are the few steps that could be followed.

Make a list of parameters, which have to be defined. Since the important parameters are fundamental categories, they have to be well defined and sufficient clues have to be provided. Better to state the conditions at different levels.

With clear-cut no ambiguous parameters and collocation, ambiguities could be eradicated to certain extent. As done in the minimalist program, a step-by-step approach could be adopted. For tagging purpose, Ranganathan's residue method could be followed. Based on the parameter a good program should be written that has to be followed by the computer for automatic tagging.

## 12. CONCLUSIONS

The minimalist program can be applied to indexing language, if it is an analytico-synthetic model. For the grammatical module in the NLP environment, X-Bar theory may be applied which is ideal for indexing language. It is suggested that the parsing be done choosing the fundamental categories as the parameter and proposing the rules relevant to the need of IL. In this model, we will have more independence to draw the choices and develop parsers. Based on this, developing UIL will not be a problem. A very important fact is that Ranganathan's facet syntax has to be made universally acceptable. For developing universal indexing language and also to apply the minimalist program CC has

to undergo a few modifications in its postulates and principles or rules on the whole. Hoping that the combination of Chomskyan and Ranganathan's paradigms would be able to contribute to the classification in the digital environment, the present study proposes a model using the Chomskyan minimalist program.

---

## REFERENCES

1. Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
2. Narasimhan, R. 1992. "Cognitive Paradigms In Knowledge Organization: A Key-Note lecture." 2nd International ISKO Conference, Madras, August 26-28.
3. Chomsky, Noam. 1992. *A Minimalistic Program for Linguistic Theory*. MIT Occasional Papers in Linguistics.
4. Sinha, Anjani Kumar. 2000. "Lectures on the Chomskyan Paradigm," Mysore: Central Institute of Indian Languages.
5. Chomsky, Noam. 2000. *The Architecture of Language*. New Delhi: Oxford University Press.
6. Neelameghan, A. 1971. Sequence Of Component Ideas in a Subject. *Library Science*, vol.8; paper Q.
7. Sharada, B. A. 1995. Infolingistics: An Interdisciplinary Study. *Library Science with a slant to Documentation and Information Studies*. Vol.32 No.3 pp. 113-121.
8. Sharada, B. A. 1997. Informetrics and Subject Indexing Language. *IASLIC Bulletin*, 42(3) September.
9. Sharada, B. A. 1998. Rules derivation for Kannada based indexing language using transformational grammar. *Library Science with a slant to documentation and Information Studies*. Vol. 35. No 22 pp. 133-138.