

ON THE CHOICE OF AN APPROPRIATE FRAMEWORK FOR COMPUTATIONAL LINGUISTIC RESEARCH ON INDIAN LANGUAGES

B.N.Patnaik
Department of Humanities and Social Sciences &
Centre for Creative Writing and Publication
Indian Institute of Technology Kanpur
Email: patnaik@iitk.ac.in

The paper is basically about the suitability of the grammatical systems developed in the west for the description of the Indian languages, for the purpose, in particular, of computational linguistic research. If we generalize things a bit, we know that it isn't really a new issue. American structural linguists were concerned with this decades ago, and so was, quite independently, the Oriya linguist and intellectual Nilakantha Das, who was working on Oriya grammar, in the early twenties of the last century. He realized that Oriya could not be best described in terms of Sanskrit grammar, and in fact, he felt that certain ideas from the grammar of English would be most useful for his purpose. Das perhaps did not make a general methodological observation in this regard, but his position would have hardly been really different from the one that each language is unique and is best described in its own terms – Das had no problems in borrowing ideas from other grammatical traditions, if he found that they contributed to a better description of the facts.

We know that the significance of a certain issue can be different in different times. Consider an aspect of the contemporary intellectual milieu. We not only live in a knowledge society, we also live in a post-colonial, post-modern, globalizing society. We are aware of the more recent thinking on the way knowledge is constructed and evaluated. In the eyes of power, traditional knowledge of the colonized people is backward looking, primitive, constrictive, and western knowledge is forward looking, modern, civilizing and progressive – a perspective that leads to a justification of colonization. Post-colonial thinking makes us intensely aware of the power motivation that plays a crucial role in knowledge creation. Post-modern thinking rejects universality in the knowledge domain and asserts that the so-called explanations are merely stories, and that no story is more privileged than any other; there is no truth to discover. For Chinua Achebe, story telling has to do with power, so only the victim can tell her own tale in an authentic manner; the stories about the victim that the victors have told lack authenticity. Turning to globalization, it is seen in certain quarters as part of the imperialistic design of the affluent and the powerful to establish their hegemony, not merely in terms of military power, but also in terms of culture. The intellectual in the so-called third world is aware of these currents of thought.

Then there is the other aspect: there has been considerable disenchantment with the western models of economic development, systems of governance, etc. It is felt, rightly

or wrongly is not the question here, that these models of development has made the wealthy wealthier, and the poor, poorer. There is the anxiety that the capitalist system that globalization would impose would make the life of the poor even more miserable. In the context of India at least it is often believed that the system of administration that India inherited from the colonial rulers has largely been rather insensitive to the needs of the people, especially the marginalized people, and it is also believed that many of the problems that the country faces today, including Kashmir and communalism, are part of the colonial legacy. Here one is talking about only the perceptions, so there is no need for debating whether all this is fair.

One reaction to the above is to explore the possibilities of the indigenous knowledge systems, which at one level may be seen as an attempt to establish one's identity and acquire and assert one's self-respect in a world that is threatening to be culturally homogeneous, where this process is seen, especially in the third world countries, to be organically related to hegemony of the west. This paper partially situates the issue of the appropriateness of western grammatical models for the Indian languages in this intellectual milieu.

There is another aspect to this issue: it is intuitively satisfactory to subscribe to the view that knowledge systems are what they are because of the facts they describe. For example, English is a fairly fixed word order language, whereas the Indian languages are fairly free word order languages. This property of the latter has to do with the relative richness of the inflections, whereas English having relatively poor inflectional morphology expresses grammatical relations, such as subject, object, etc. in terms of fairly fixed word order, using informal terminology. One could observe, using theory neutral terminology, that the grammars of Indian language would have very little by way of syntax, unlike the grammar of English, i.e., these are essentially different systems, which conclusion is reinforced by certain other facts as well. Reduplicatives, compound verbs, serial verbs in some languages, such as Tamil, etc. are not encountered in English, whereas auxiliary *do*, question word in the initial position of a clause, subject-auxiliary verb inversion, introductory *it* and *there*, etc. do not have equivalents in Indian languages. Therefore it sounds quite persuasive that the model for English would be inappropriate for the description of Indian languages (actually morphologically rich languages, but here the preoccupation is with Indian languages, which is why we do not wish to generalize beyond the Indian languages.). This way of looking at the questions under discussion gains a great deal of credibility and support in the context of the intellectual milieu outlined above.

Exploration of the possibilities of the indigenous knowledge systems in the Indian context in the case of language leads one to Panini and the later grammarians including Bhrtrhari, the *nyayayikas*, the *mimansakas*, the *alankarikas*, Tolkapiyam, etc. We confine ourselves here to the grammarians: Panini and Tolkapiyam, and we use the term ancient Indian grammatical tradition to refer to the seminal ideas in both, which are almost the same. The tradition was concerned with offering a systematic and an economical description of the language under study, say, Sanskrit, and also with the question of how language is learnt. On the latter, the ancients' views are very similar to those of the

behaviourists: essentially, language is learnt from observation, experience, and through practice. Teaching plays a very significant role in it, and it is through this that the two concerns – language structure and language learning - are connected; the language can be best learnt with the help of the best description of the language – it is possible obviously only in terms of the best grammar of the language. For the best grammar of the language, the data comprise a selection of the languages used, namely, the language used by the *sistas* (roughly, the elite). Thus in a sense the grammar is “corpus based”, if “corpus” is understood in a somewhat more extended sense: it does not refer to any actual corpus, but a hypothetical inventory of a selection of all well formed linguistic forms of the language. Being word order free clause internally, and being morphologically rich, the language under reference is most economically describable within the possibilities of a “lexicon centred grammar” in some sense. The grammar thus is essentially a word grammar, dealing at great length and in great precision the derivation of the words from their roots and affixes. In a clause the words satisfy the mutual expectancy requirement, and the relations between them are, basically, morphologically realized. So this is the nature of grammar, it is believed in the intellectual milieu outlined above, which is most likely to prove to be the best candidate for the purpose of the description of the modern Indian languages, since it is generally held that these are free word order languages and are morphologically rich, as is Sanskrit. The high commendation that *Astadhayi* has received and is still receiving from the western linguists has contributed in a considerable measure, to the confidence of the scholars who subscribe to the view that the modern Indian languages are best describable within the Paninian (which in the present paper also includes Tolkapiyam) framework.

During the colonial period Indian languages came into contact with English and European scholarship came into contact with ancient Indian scholarship, as did the Indian scholars with European knowledge. Recognition of the knowledge created in India in the west and Indian renaissance, which drew inspiration from the west were among the affirmative consequences of this interaction. The grammars of Indian languages, the grammar of Oriya, to consider a specific case, taking Oriya as representative of modern major Indian languages, were influenced by the available grammars of English. The notions of subject and object were incorporated into the grammar of Oriya (as of other languages). These are not notions of ancient Indian grammatical scholarship; although *karta* and *karma* are often considered to be equivalents of subject and object, these indeed are very different notions. In the Paninian system, *karta* is characterized as “independent”, and *karma*, “the most desired by *karta*”, subject controls verb agreement, whereas in the Indian system *karta* does not control the verb morphology – in fact, the idea of subject-verb agreement itself comes from English grammars, as is the notion of eight parts of speech, subject-predicate division of a sentence, direct and indirect speech distinction, to name a few. In sum, the grammar of Oriya incorporated decades ago notions from English grammar in order to increase its descriptive power. Nilakantha Das strongly supported and practiced this eclecticism in his grammar.

Interaction of the Indian languages with English resulted in the borrowing of certain constructions from English, such as the indirect speech construction, the periphrastic passive construction, the so-called English-type relative clause construction, etc.,

confining ourselves to borrowing at the level of syntax. These are at different stages of nativization, as far as Oriya is concerned. New constructions have also arisen as a result of this contact, such as the indirect speech construction that uses the complementizer (corresponding to English *that*) that separates the quoted material from the main clause, and the direct speech material in the embedded clause:

- (1) *rama kahila je mu gharaku jibi*
(Ram – said – COMP – I – home to – will go)
(Ram said that I (=Ram) will go home.)

Notice that in this indirect speech construction, the subordinate clause is in the direct speech form, with no pronominal change, etc. Consider the one borrowed:

- (2) *rama kahila je se gharaku jiba*
(Ram – said – COMP – he – home to – will go)
(Ram said that he (=Ram) will go home.)

Notice the use of the complementizer, and the pronominal change in the subordinate clause. Notice also that the matrix clause and the subordinate clause do not have the same tense; the so-called “sequence of tense” phenomenon does not exist in Oriya. For the present purposes it may be noted that the borrowed construction does not put any pressure on the morphological system of the recipient language. If the passive construction (3) went out of existence, it was because it demanded a modification at the morphological level:

- (3) *eha kari para jiba*
(this – do non finite – can passive – go AGR)
(This can be done.)

According to Nilakantha Das this construction existed in Oriya and obviously the structure was borrowed from English. But it disappeared because it demanded the creation of the passive participial form of the defective modal verb *par*. Thus one could surmise that nativization of a borrowed construction is possible when it does not lead to the modification in the morphological system of the language, especially those involving the functional words of the language. This shows that the borrowed constructions are describable within the possibilities of a lexicon centred system of the Paninian type.

This is not to say that no augmentation is necessary. For instance, consider the category of the so-called indeclinables. Traditional grammars of Oriya arrange these in subsets on semantic and pragmatic considerations, but a more detailed arrangement is needed. In addition, their syntactic subcategorization is also necessary. Consider *ti* (and *ta*) for example: it is an attitudinal term, but *ti* can be used as a definitizer, a numeral, a classifier, an emphazier, etc. The study of adjuncts is very inadequate in traditional grammar, not only of Oriya (and major modern Indian languages of which, it may be recalled, Oriya is taken as representative for the purposes of the present paper). Adjuncts

constitute a neglected area of traditional grammar. Their syntax, semantic and pragmatic properties await detailed and careful study.

The suitability of a model for a language is in the ultimate analysis a question of the objectives of the study of that language. Therefore whatever be the intellectual milieu, or the inherent strength of a particular indigenous knowledge system, if the objectives of the given project are not served in at least some satisfactory manner, then the search has to be for a non-indigenous model. If for developing language technologies the traditional grammatical knowledge is in principle inadequate or very uneconomical, then choice has to be made in favour of such knowledge created elsewhere. On this consideration it is not clear that the Indian traditional grammars are unsuitable for the purpose, although it is an empirical question. For the specific purpose, what are needed are good corpuses, and the grammar must be able to describe the linguistic forms in the corpus. If the language has the property of morphological richness, then the grammar must contain a powerful system for morphological analysis. It must have a rich repertoire of labels with which to provide a detailed tag to each unit of the linguistic form at various levels of analysis. As already observed, it is not clear that the Indian traditional grammars are *prima facie* inadequate.