Computational treatment of Indian languages: problematics

Gérard Huet

Inria Paris Center,
France

Abstract

Indian grammatical tradition [*Vyākaraṇa*] predates Western linguistics by 25 centuries. Many central notions of universal linguistics, such as phoneme [*varṇa*], stem [*pratipādika*], inflected word [*pada*], semantic role [*kāraka*], dependency [*ākāṅkṣā*] were already known from Pāṇini as explicit concepts. The notion of *sign*, combining an utterance with its meaning, is implicit from the formulation of his grammar. Thus morphemes such as *kṛt* suffixes operate both on the phonemic stream of an utterance and on its meaning, through generic paraphrases. Operations are detailed to the point of being similar to computer instructions, and thus correct Sanskrit generation can be reduced to writing a computer emulation of the rules of Aṣṭadyāyī.

Of course correct enunciation is only one side of the coin, the side of the speaker. If we model his communication intention as a sequence of *sūtra* invocations, we may more or less use deterministic computing to produce the prose order of an enunciation with intended meaning. The other side of the coin, understanding by the speaker [*śabdabodha*], is much harder to model. Specially if we parse text, since classical Sanskrit is written without accents, so the prosody is not available. Further, poets used the relatively free order nature of the language to write complex sentences with dislocated prose, making it a tough problem to guess plausible prose orders amenable to structural decomposition. Finally, the smoothing of the voice signal by sandhi makes an additional task, segmentation [*sandhiviccheda*], necessary. This not only induces non-determinism, it induces ambiguities that may be intended by the speaker as double-entendre [*śleṣa*].

For all these reasons, understanding Sanskrit by computer is a tough task, even though we have a perfect grammar for the language. Understanding a Sanskrit sentence is not just a matter of parsing a paraphrase, it involves recognizing compatibility in meanings [*yogyatā*] of components of the sentence, involving the choice of possible usages of words through the nature of their acceptions, as primary etymological sense [*abhidhā*], figurative/metaphoric usage [*lakṣaṇā*] and allusive meaning [*vyañjana*]. Understanding such processes goes beyond mere *Vyākaraṇa*. It involves philosophical questions concerning the nature of reality and its relationship to language, as well as pragmatics of communication, epistemic concerns such as reliability of knowledge acquisition [*pramāna*], and finally esthetics concerns like emotional response [*rasa*] in literary theory [*saahitya*].

To account for these aspects of *śabdabodha*, the Indian tradition developed a lot of conceptual material under the various philosophical points of view [*darṣaṇa*]: *nyāya*, *vaiśeṣika*, *mīmāṃsā*. It is not easy to relate philosophical concepts across traditions with different ontological classifications. For instance, modern mathematical logic does not correspond to *nyāya*, but rather to *tarka*. The philosophical realism of *nyāya* mixes arguments of physical knowledge, even if they appear naive in an anachronistic comparison with modern science, with deductions through a notion of pervasion *vyāpti*. This lead to the development of a conceptual calculus within *navyanyāya*, where a controled use of Sanskrit compounds led to unambiguous semantic descriptions. In parallel, *mīmāṃsā* developed notions of pragmatics such as economy and focus of discourse. But this also blended with an esthetic movement originating from dramaturgy [*nāṭya*] which further

developed in literary theory with Ānandavardhana's notion of suggestion [*dhvani*]. This was further pursued with Abhinavagupta's school within Kashmir Shivaism.

In the West, linguistics really started only in the 19th century, with the pioneer Swiss scholar de Saussure, who actually got a start on the problematics of language with the study of Sanskrit. More recently, the advent of computers gave a new impetus to the discipline with the development of mathematical and computational linguistics. The semantic modeling of language has been an important component of the field of artificial intelligence, with topics such as knowledge representation, common sense reasoning, belief revision, sentiment analysis, etc. Some of this research reinvented *vaiśeṣika* ontological classifications into endless variations, and developed soft concepts such as fuzzy reasoning and non-monotonic logic which did not lead to convincing realisations. On the other hand, computational logic led to successful efforts at the mechanisation of mathematics and the certification of computer programs according to logic specifications. Formalisms blending type theory with modal logic have been proposed to develop knowledge structures adequate to model discourse, such as Montague semantics, discourse representation theory, Barwise's situation logic, hybrid logic, etc. We must also mention efforts at designing semantics-informed lexicons, such as Wordnet, and lexical vector representation models.

In India, computational linguistics used both Western methods for vernacular languages, and *Vyākaraṇa* concepts, mostly for Sanskrit. A plea for using traditional knowledge systems for artificial intelligence problems was voiced as early as 1984 by Rick Briggs, but this has not been followed by practical developments so far. Attempts have been made to adapt Western computational linguistics notions such as Wordnet to Hindi and Sanskrit. Paninian methods have been suggested as practical solutions for the treatment of natural language in general, such as the Akshar Bharati movement and the subsequent Saṃsādhanī software of Amba Kukarni. The use of such methods seems to be specially promising for Indian vernacular languages, since common cultural notions may be shared through multilingual lexicons, name-entity relationships databases, semiotics thesauri, etc. Also the *navyanyāya* concept calculus, almost unknown in the West, ought to be studied in comparison with calculi developed in mathematical logic and formal semantics notations.

The whole area of computational linguistics is undergoing a shift of paradigm towards corpus linguistics methods based on statistical techniques (big data, deep learning, etc.) Use of this new computational methodology makes sense, now that enormous amounts of linguistic data are available through Internet. However, its practical application to Indian languages is still to be demonstrated, in view of the slow development of appropriately tagged linguistic data to be used as training corpus, and on the specificity of the cultural area. We are now at the crossroads of many possible paths of investigation for the computational treatment of Indian languages.

Actually, computational linguistics for Indian languages covers many orthogonal aspects, with respect to the applications aimed at. One area of application is supplying Indian citizens with the means of accessing essential services through numerical interfaces such as the Web in their local vernacular language. Another one, quite different, would be to develop cultural heritage tools helping to preserve resources such as litterature across all languages over the global Indian cultural area.