Towards Computational Processing of Sanskrit

Gérard Huet

INRIA

ICON-2003, December 20st 2003

Sanskrit milestones

- Indo-European family
- Indo-Iranian branch
- Vedic
- Pāṇini
- Classical Sanskrit
- The Sanskrit Corpus
- The *paṇḍita* tradition
- Prakrit, Pāli, Hindi, Bengali, etc

Sanskrit Indian Tradition

- Pāṇini's generative grammar (*aṣṭādhyayī*, *śivasūtras*, *dhātupāṭha*, *gaṇapāṭha*)
- Katyāyana, Patañjali
- Bhartṛhari's *vākyapadīya*
- Sāyaṇa's commentary of Veda
- Vedāṅga - *śikṣā*, *vyākaraṇa*, *chandas*, *nirukta*

Comparative linguistics and Veda

- Wilson Roth Müller
- Bergaigne Oldenberg Bloomfield
- Whitney Apte
- Böhtlingk Monier-Williams
- Speijer
- Renou Gonda

Modern descriptive linguistics

- Cardona
- Kiparsky
- Hock
- Deshpande
- Aklujkar
- Scharf
- Gillon

What is the syntax of Sanskrit ?

(Government and Binding revisited)

- 1st approximation: free order within VP
- Staal's Constraint: the Calder mobile theory
- Discontinuous constituents analysis (crossings)
- 3 kinds of dislocations/extrapositions
- Absolutives as sentinels
- Sharing of agent between gerund and main clause
- Null arguments and anaphora for extra sharing

Control in Sanskrit: Bhartṛhari's rule

Analysis from Brendan Gillon.

(1) **कष्ठानि भित्त्वा ओदनम् पचति देवदत्तः**
kaṣṭhāni bhittvā odanam pacati Devadattaḥ
Having split wood, Devadatta is cooking rice.

(2) **पक्का ओदनम् भुङ्क्ते देवदत्तः**
paktvā odanam bhuṅkte Devadattaḥ
Having cooked the rice, Devadatta is eating it.

Here the patient (*karman*) is shared ! LL Contraction. What about sharing of items with distinct *kāraka* (semantic role) ?

Bhartṛhari's rule. The main action determines the *kāraka*, hence the explicit case. The subordinate *kāraka* is implicit. This was explained by Deshpande.

Such rules are essential to understand valence constraints.

Why don't you just implement Pāṇini ?

Firstly, Pāṇini is complex, and, as we saw, not sufficient. However, it is probably one of the most accurate formal description of any human language, and many criticisms of Pāṇini arose from an imperfect comprehension of his framework.

Secondly, the primary feature of his formal presentation is conciseness (*lāghava*). This conciseness sometimes is counter-productive, because the abbreviation principles may be without linguistic meaning, as discussed by Kiparsky.

Finally, a generative grammar, even if perfect, does not provide the means to analyse language unambiguosly. Pāṇini's description proceeds through the linguistics layers from semantic information to morphosyntactic representation to abstract morphological form to phonological form. Information from upper layers is available at lower layers, but not conversely.

Pāṇini ought to be fully computerised

Still, it is useful to compile Pāṇini's *aṣṭādhyayī* into a computerised form. This is being done e.g. by Dr Shivamurthy Swamiji in Sirigere with his *gaṇakāṣṭādhyayī*.

Visit `http://www.taralabalu.org`.

NB. The system lists the full trace of rewritings in terms of sūtras.

Going in the reverse direction

- Phonology (Sandhi)
- Morphology
- Root Lexicon
- Flexed Lexicon (invertible)
- Segmentation Automaton
- Tagging Transducer
- Constituants Analysis
- Agreement/Valency
- Anaphora resolution

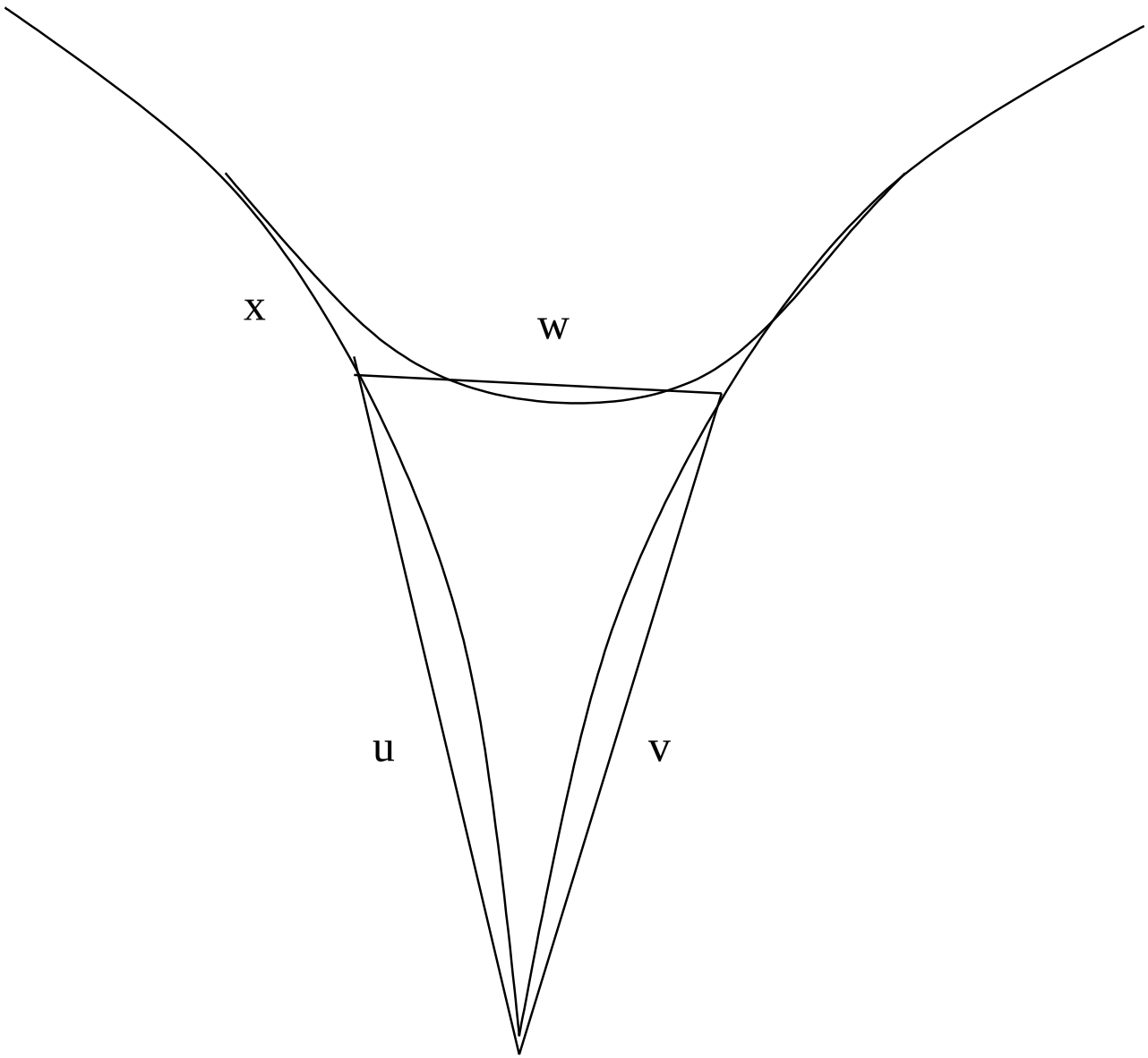
Euphony and sandhi

When successive words are uttered, minimisation of the energy necessary to reconfigurate the vocal organs at the juncture of words leads to a phonetic transformation, discretized at the phonemic level by a *sandhi* rewrite rule of the form:

$$[x]u|v \rightarrow w$$

This juncture euphony, called external *sandhi*, is explicit in sanskrit in the written form of the sentence, where words are merged together in a continuous phoneme stream. A similar operation glues compound words together.

The first computer treatment of sanskrit is thus segmentation analysis, a complex non-deterministic process.


x
w
u
v

Variety of sandhi

- Internal sandhi: frozen flexed forms
- External sandhi : compound words
- Terminal sandhi
- Morphological sandhi
- Syntactic sandhi
- Phantom phonemes

Exemples

- $tad + śrutvā = tacchrutvā$
- $dviṭ + hasati = dviḍḍhasati$
- $vāk + me = vāṅme$
- $a + i = e$
- $a\|ā + a\|ā = ā$
- $dṛś +\!+ ta = dṛṣṭa$
- $ā +\mathrm{m}\ ihi = !ehi$
- $iha +\mathrm{s}\ !ehi = ihehi$
- $iha +\mathrm{s}\ ihi = ihehi$
- $iha + ehi = {}^*ihaihi$

General methodology

- Dictionary
- Root lexicon
- Internal sandhi generator
- Morphological paradigms
- Invertible flexed forms lexicon
- Segmentation automaton

x
u
z
u
w
v
v

Properties

- Correctness
- Completeness
- Termination
- Non sensitivity to preference ordering
- Training

Problems

- Non-determinism
- Overgeneration
- Bahuvrīhi
- Dual forms, *sa*, etc
- Reference corpus
- Training
- Robust mode
- Lemmatisation
- Lexicon acquisition

Syntax

- Agreement, chunks, genitives
- Valence, sub-categorisation
- Role features constraints
- Sentinels
- Sharing, dislocation
- Double accusatives, etc
- Minimal ontology
- Anaphoras, ellipses
- Problems with *iti* etc.

Semantics

- Proper names, titles, surnames
- Metaphors
- Myths
- Sūtras, commentaries, tradition
- Esoteric language, mantras, poetry
- Cultural heritage

Available Resources in India

- CDAC (and others) Fonts
- CDAC Ramanujan's Desika, Sanskrit Authoring system
- Bhandarkar Institute - Mahabhārata Critical Edition
- Deccan College Sanskrit Dictionary Project
- Academy of Sanskrit, Melkote. Shabdabodha
- Jawahar Lal Nerhu Univ. CASTLE software, Sanskrit tutor
- Swamiji's *Gaṇakāṣṭādhyayī*, Sirigere
- Tirupati's Sanskrit Net corpus
- Possibly many other efforts...

Available Resources in the West

- Dictionaries: Köln, etc
- Numerous Web sites with various documents
- Corpus: Takunaga, Smith, Bhandarkar
- Peter Scharf's Sanskrit Library at Brown
- Sanskrit Reader for *Rāmopākhyāna*
- Whitney's Roots
- Gillon's Apte compilation
- Towards a Sanskrit Treebank

Gillon's analysis of Apte's corpus

Extracted from "Word order in classical Sanskrit", by Brendan Gillon. The sentence, a real quotation from *Mudrārākṣasam*, is listed in Apte's "Student Guide to Sanskrit Composition". It is an example of extraposition to the right periphery of a clause.

(15) Mu 2.10.17 ⟨= SG 12.1.2 ⟩

[$_S$	[$_{ADV}$	na	]	[$_{NP1s}$	[$_{NP6}$	naḥ	]	kutūhalam	__	]	[$_{VP}$ asti]
		not				our		curiosity			is

[$_{NP7}$	(sarpa	-	darśane)	]	]	
	snake	-	seeing			

There is no curiosity on our part to see the snake.

Synthesis of the hierarchical structure

Conjecture. It should be possible to synthesise the constituents structure from the linear structure by computing valence and agreement constraints satisfaction modulo movement. Minimalism ?

But it is expected that nonlinearities will raise complexity issues.

Relevant Conferences

- Sanskrit International Conference
- SALA round table workshop
- Workshop on Sanskrit Informatics Delhi Aug. 03
- ICON

Towards a Sanskrit Perseus?

- http://www.perseus.tufts.edu/
- Cooperation on normalised SKT Resources
- Sharing of corpus data
- Interoperability of tools
- Evaluation
- Sustainability of long term effort

Zen technology

- Lexical trees (tries)
- Zippers/contexts
- Sharing Functor
- Minimal automata
- Difference words
- Lexicon morphisms
- Automata mista

Enjoy!

- Sanskrit site: `http://pauillac.inria.fr/~huet/SKT/`
- Sandhi Analysis paper:
  `http://pauillac.inria.fr/~huet/FREE/tagger.ps`
- Course notes:
  `http://pauillac.inria.fr/~huet/ZEN/esslli.ps`
- Tutorial slides:
  `http://pauillac.inria.fr/~huet/ZEN/Hyderabad.ps`
- ZEN library: `http://pauillac.inria.fr/~huet/ZEN/zen.tar`
- Automata mista:
  `http://pauillac.inria.fr/~huet/PUBLIC/zohar.pdf`
- Objective Caml: `http://caml.inria.fr/ocaml/`