Anusāraka: Machine Translation and Language Accessor

Akṣara Bhārati

The International Conference on the Contribution of Advaita Vedanta to Humanity

Nov 21, 2015
Anusāraka: An effort towards addressing the problem of reducing language barriers in multilingual context such as India.

Problem is Large, Complex and Highly Challenging
There is no loss of effort
nor is there any harm
(of production of contrary results).
Even a little of this knowledge,
even a little of this yoga,
protects one from the great fear.
Brief History

Founded by Br. Vineet Chaitanya in 1984

Inspired by Gurudev

- East meets West
- Bridge the gap between traditional knowledge and technology
- Pundit and Public
Personification of a group working on

- Computational processing of Indian languages
- Giving due importance of the traditional Indian theories of language
- Team work (leading to personification)
Spirit of Akṣara Bhārati

- **Bourbaki (Nicholas Bourbaki)**
  - Pseudonym for a group of, mainly, French mathematicians, starting in 1935
  - Wrote a series of books presenting an exposition of modern advanced mathematics
  - Rigour and generality
**Durgā (Mahishāsura Mardini)**
Created by giving their best by

- Śivā (the destroyer) - the trident
- Viṣṇu (the Protector) - the conch
- Agni - Agni (god of fire) - the spear
- Yama (god of death) - the cudgel
- Vāyu (god of wind) - the bow
- Sūrya (sun) - the arrows
- Indra (god of rain) - the vajra
- Kubera (god of wealth) - the mace
- Brahmā (The Creator) - the water pot
- Kāla (Time) - the sword
- Viśvakarma (god of architecture) - the axe
- Himavān (mountain god) - a mountain lion as her vehicle
Journey started at IIT Kanpur
  - Starting with Sanskrit
  - Later among Modern Indian Languages

Connected with many institutions
Rashtriya Sanskrit Vidyapeetha, Tirupati
University of Hyderabad
IIT Mumbai, etc.

Reached IIIT Hyderabad (1998)
started work on English-Hindi
Anusāraka: Different dimensions

- A practical demonstration of application of traditional śāstras to solve contemporary problems
- A tool to overcome language barrier
- A better approach for building Machine Translation Systems
- Language Teaching through Applied Grammar
- An opportunity for masses to be IT contributors rather than mere IT consumers
Claim:
Pāṇini was aware of the strength of language as an information coding device.
And Pāṇini made the best use of this strength.

Evident from

- His style of presenting the information in sūtra style, and
- The way he has analysed the Sanskrit Language
Māheśvarasūtra

⇒ Importance of information coding

Syntactico-semantic Analysis

⇒ Information encoding: Some insights
  - Where
  - How much does a language code the information
  - How
Pāṇini: An Information Scientist

1.

Māheśvarasūtra

a i u \(\mathbb{N}\)
\(\check{r} \mid K\)
e o \(\check{N}\)
ai ao C
h y v r T
l \(\mathbb{N}\)

\(aṇ \Rightarrow \{a \ i \ u\}\); \(aṇ \Rightarrow \{a \ i \ u \check{r} \mid e \ o \ a i \ a o \ h \ y \ v \ r \ t \ l\}\)

\(iṇ \Rightarrow \{i \ u\}\); \(iṇ \Rightarrow \{i \ u \check{r} \mid e \ o \ a i \ a o \ h \ y \ v \ r \ t \ l\}\)
Mahābhāṣya on the apparent ambiguity

5 sūtras with a\N

ढू लोपे पूर्वस्य दीर्घः अणः ६.३.११०
के अणः अजस्य हस्वः ७.४.१३
अणः अप्रगृहस्य अनुनासिकः (वा) ८.४.५६
उरण् रपरः १.१.५०
अणुदित सवर्णस्य अप्रत्ययः १.१.६८
सामर्थ्यः (Ability to convey proper meaning)

ढ़ लोपे पूर्वस्य दीर्घः अणः 6.3.110
के अणः अक्ष्य हस्वः 7.4.13
अणः अ-प्रगृहस्य अनुनासिकः (वा) 8.4.56
हस्वः and दीर्घः properties of a vowel.
Only Vowels can get प्रगृहस्य सञ्ज्ञा
प्रसिद्धि (Frequency of usage)

उरण्णरः १.१.५०
No example involving members of bigger set.

- The effect of the rule is nullified by other sūtra, OR
- The application of sūtra leads to undesirable redundancy in some other sūtra
किञ्चि (indicator/marker)

अणुदित सवर्णस्य च अप्रत्ययः
उःऋत् (== > तपर)
तपरः तत्कालस्य (सवर्णस्य )
== > सूत्र is applicable for ‘ॠ’
and ॠ € aN_2
== > ण is the second ण
लाघव  (economy)

इ उ == > य्व
इण: == > य्वोः
1+.5+1+.5 (=3) .5+.5+2+.5 (=3.5)

व्याख्यानतो विशेष प्रतिपत्ति: न हि सन्देहात् अलक्षणम्
Had Pāṇini used some other consonant as an anubandha, he would have lost an opportunity to train the students in paying attention to the different means of information coding a language employs.
Should we then not conclude that

Pāṇini was aware of ambiguities a natural language has and wanted to train the students of vyākaraṇa to pay attention to different sources of information available for disambiguation? And that he uses the very first opportunity to train the students – right from the Māheṣvarasūtras with which the study of Aṣṭādhyāyī commences?
II. Dynamics of Information coding in Sanskrit

- Where
- How much does a language code the information
- How
‘Where’ is the information coded?

राम: ग्रामम् गच्छति
रामेण ग्राम: गम्यते
Where is the information is coded?

First reaction:
If kartari prayoga(active voice)
  - kartā – > Nominative Case
  - karma – > Accusative Case
If karmaṇi prayoga(passive voice)
  - kartā – > Instrumental Case
  - karma – > Nominative Case
Where is the information is coded?

It is also necessary to

- state noun-verb agreement
- account for pro-drop as in gacchāmi
Where is the information is coded?

- कर्मणि च भावे च अकर्मकेभ्यः (कर्तरि) 3.4.69
- अनभिहिते 3.1.1
- कर्तृ-करणयोः तृतीया 2.3.18
- कर्मणि द्वितीया 2.3.2
- प्रातिपदिकार्थलिङ्गपरिमाणवचनमात्रे प्रथमा 2.3.46
Where is the information is coded?

If it were nominal-suffix
how to account for the pro-drop case
gacchāmi?

It is verbal-suffix which marks the relation.

What about the other relata?

Pro-drop only in case of
First and Second person pronouns.

लः कर्मणि च भावे च अकर्म्मेवथः (कर्तिरि) 3.4.69
Then what does nominative case signify?

प्रातिपदिकार्थिलिपिपरिमाणवचनमात्रे प्रथमा 2.3.46
Where is the information is coded?

Other relations:

- अनभिहिते 3.1.1
- कर्तृ-करणयोः तृतीया 2.3.18
- कर्मणि द्वितीया 2.3.2
Dynamics of Information Coding

- ठ: कर्मणि च भावे च अकर्मकेभ्यः (kartari) 3.4.69
- अनबिहिते 3.1.1
- कर्तृ-करणेऽः तृतीया 2.3.18
- कर्मणि द्वितीया 2.3.2
- प्रातिपदिकार्थैवः चपरिमाणवचनमात्रे प्रथम 2.3.46
"How much" information is coded?

विभक्ति: = f(कारक:, प्रयोग:)

1. राम: कुशिकया तालम् उद्दाटयति
2. कुशिका तालम् उद्दाटयति
3. ताल: उद्दाट्यते
How much information is coded?

राम → Agent
कुछिका → Instrument
ताळः → Patient

रामः कुछिका ताळः → कर्ता

स्वतन्त्रः कर्ताः
Greatness of Pāṇini lies in identifying **EXACTLY HOW MUCH** information is coded in a language string.

⇒

Upper Bound for the possible Analysis using only a language string and grammar.
We can extract only that which is available in the language string ‘without any requirement of additional knowledge’.

Analogy:
We can not do high quality work with low quality energy.
How is the information coded?

Bhartṛhari in Vākyapadīyam states (3.7.81-82),

प्रधानेतयोर्यत्र द्रव्यस्य क्रिययोः पृथक्
शक्तिगुणाश्रया तत्र प्रधानमनुरूध्यते
प्रधानविषया शक्ति: प्रत्येनाभिधीयते
यदा गुणे तदा तददनुक्तापि प्रकाशते

vibhakti → only one kāraka

रामः दुग्धम् पीत्वा शालाम् गच्छति

2 verbs with 2 expectancies each, and only 3 nouns!
Who drank milk?

समानकर्तर्कयोः पूर्वकाले

रामः दुग्धम् पीत्वा शालाम् गच्छति
Information is coded as a "Language Convention"
→ Different Languages may have different conventions.
→ Automatic Translation may lead to ungrammatical sentences
वनात् ग्रामम् अध्य उपेत्य ओद्नम् आश्वपतेन अपाचि.
* Having reached the village today the rice has been cooked by Aśvapata.
Where does the language code information?
   Useful to decide the parsing strategy

How much information does it code?
   Useful to decide whether the information can be passed on to
   other language without any special efforts or not

How does a language code information?
   Useful to decide whether the desired information can be
   extracted merely from a language string or not

Claim: Any grammar that is developed with these questions in
   mind will be a grammar truly in Pāṇinian Spirit.