Semi-automatic analysis of Navyanyāya compounds

S. R. Arjuna & Gérard Huet

Sanskrit Studies UoH & Inria Paris-Rocquencourt

30th SALA
February 8th, 2014
University of Hyderabad

Navyanyāya style

- Long compounds
- Specific technical vocabulary
- Productive use of various *taddhita* suffixes
- Semi-formal compound structure
- Possible discrepancy with Pāṇinian rules ?

Collection of *Navyanyāya* compounds

S. R. Arjuna has been studying work by Gaṅgeṣa on the concept of *vyāpti*. This logician gives a set of 5 definitions, called *vyāptipañcakam* or *pañcalakṣaṇī*. He then sets to refute them, and quotes two previous definitions, by Tārkikasimha and Tārkikavyāghra, together known as *Simhavyāghralakṣanam*. This work is commented in a book "*Vyāptipañcakam simhavyāghralakṣaṇam ca*", which was selected by Arjuna as typical corpus of *Navyanyāya* discussions.
He extracted from this corpus a set of 14,000 exemples, mostly nominal compound examples. From this digital corpus he chose at random 356 examples for the purposes of the present experiment, aiming at using the Heritage Reader for semi-automated processing of *Navyanyāya* corpus.

Experimental version of Heritage Reader

At mid-december 2013, after learning of Arjuna's difficulties, Gérard set to design an experimental version of the Heritage Reader, tuned for recognition of *Navyanyāya* compounds.
Here are the salient features of this extension:

- Databanks are added for *taddhita* suffixes inflected forms
- Suffixes involved are *-taa* (Fem), *-tva* (Neu), and *-vat*/*-vatī* (all 3 genders)
- Segmenter transitions added to accommodate *taddhita* productivity
- Word mode for single *pada* in order to curb overgeneration
- Graph interface used to display the huge solution space
- Technical vocabulary of *Navyanyāya* acquired in lexicon

Experiments unveiled a few weak spots to fix, such as *avagraha* treatment, and required lexicon acquisition of a small number of vocabulary items.

Experiments iteration

In mid-december 2013, S. R. Arjuna experimented with a limited set of 356 examples, and reported problems.
In january 2014, G. Huet fixed a few issues on the basis of this report, and incorporated the experimental mode in a new V2.80 version of the Heritage Engine. He then classified the remaining problems as follows.

Classification of problems on a 356 random examples selection

Firstly corpus data anomalies:

- 2 encoding problems (vowel l)
- 7 examples with *-iti* or *na-* segments
- 3 typos in corpus

These problems were solved by data cleaning/standardization.
The 7 non-compound examples were discarded from the test set.

Classification of problems (2)

Secondly processing failures:

- 3 time-out failures
- 3 -*ka* suffix
- 7 -*tā*-*ka* constructions
- 13 inchoative compounds (-*cvi*) used as component (iic. or ifc.) of an englobing compound

This gives a recall of 91%. The last category points out an incompleteness in the current compound recognizer, which deserves correction. When the corresponding extension is available, the recall ought to improve to 95%.
If the (compound) suffix -*tāka* is added as productive, the recall might reach 97%. Of course this might degrade the precision, and even may induce more silence by time-out.

Conclusion

It is possible to rapidly tailor specific versions of our software to cater to specific corpus characteristics. For the case in point, il appears that most long compounds of *Navyanyāya* texts are amenable to semi-automatic segmentation.

The correctness of the method relies on an important property of the selected productive *taddhita* suffixes, namely that the sandhi used for their (fake) declension does not overlap with the sandhi of their own affixing.

It is to be expected that the treatment of the extended set of 14,000 exemples will reveal other difficulties. Some further *taddhita* suffixes, such as *-mat*, may be required. This will not be too problematic hopefully. The general problem of making full *taddhita* recognition raises interesting issues.

An interesting conclusion is that so far the considered *Navyanyāya* corpus is basically conformant to Pāṇini.

Thank you for your attention!