Activity report 2010

Sanskrit Associated Team

Centre de Paris-Rocquencourt, INRIA (France),

Dept. of Sanskrit Studies, Univ. of Hyderabad (India)

and the Sanskrit Library, Providence (USA)

1 Summary of joint activities

This year marked a change of pace in the joint team.

After 3 years of informal cooperation between the two sites of Rocquencourt and Hyderabad, the joint team was officialised as a bilateral collaboration on Sanskrit Computational Linguistics between INRIA and University of Hyderabad. In March 2010 a Memorandum of Understanding was signed by UoH Vice Chancellor and INRIA's Chairman. It is available on the Web site of the joint team, at `http://yquem.inria.fr/~huet/EA/`. The UoH contributes to the budget of the joint team, in particular for hosting visiting guests at the Lake View UoH guesthouse and providing various facilities for joint work.

Then in September, Dr Peter Scharf, Senior Lecturer in Sanskrit at Brown University and Director of the Sanskrit Library project, who had been associated with the activities of the joint team for a long time, and organized the 2nd Symposium in 2008, decided to join the team as a third partner. Dr Scharf has a long distinguished record of Sanskrit computer-assisted philology, in cooperation with German colleagues from Max Planck Institute Berlin and Köln University. In July he won an important award of the US National Endowment for the Humanities (NEH) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in the framework of a call for international digital humanities projects (`http://archiv.twoday.net/stories/6482736/`). He will receive a grant of $177,872 to establish a digital Sanskrit lexical reference in partnership with his colleagues of the Cologne Digital Sanskrit Lexicon (CDSL) project of the Institute of Indology and Tamil Studies at Köln University. He is currently principal investigator of a 3-year grant for $301,540 of the National Endowment for the Humanities, Enhancing Access to Primary Cultural Heritage Materials of India: Integrating images of literary sources with machine-readable texts, lexical resources, linguistic software, and the web (PW50408).

Benoît Razet, who was working on his PhD research under the direction of Gérard Huet at the INRIA Paris-Rocquencourt Center, graduated at the end of November 2009 at University Paris-Diderot. His thesis, "Machines d'Eilenberg Effectives" gives a general theory of relational programming as an effective implementation of the X-machines of Samuel Eilenberg (1974). The thesis is avail-

able for download at his home page http://www.tcs.tifr.res.in/~razet/. This new foundation of relational programming actually emerged from the work on Sanskrit parsing, since it is this application (more precisely, sandhi splitting) which motivated the development of the reactive engine, the core simulator of the effective Eilenberg machines.

On January 1st, 2010 Benoît Razet moved to the prestigious Tata Institute of Fundamental Research (TIFR) in Mumbai, as a Visiting fellow. There, primarily, he continues his research on automata theory within the working group "Regularity Rationality". But he also collaborates with Amba Kulkarni, Chairperson of the Department of Sanskrit Studies, Hyderabad University and Indian Head of the Sanskrit Associated Team. At the occasion of the International Congress of Mathematicians in Hyderabad in August, they worked for a week on the problem of incorporating quantifiers in a formal abstraction of the concepts of the Navanyāya school of Indian logic.

Amba Kulkarni submitted her dissertation "Anusāraka: An approach for MT taking insights from the Indian Grammatical Tradition" at University of Hyderabad and obtained her PhD degree in August 2010. The thesis is available for download at her home page http://sanskrit.uohyd.ernet.in/faculty/amba/.

Gérard Huet continued his work on the Sanskrit Heritage platform. In September he released version 245 of the accompanying web services, at http://sanskrit.inria.fr/. The segmenter now takes advantage of all the participal stems (*kṛdantas*) generated by the morphological engine. He generalized the compound recognizer by allowing the so-called *nañtatpuruṣa* (negative compounds obtained with the privative prefixes *a* and *an*). This gives rise to an Eilenberg machine with 21 phases, whose underlying automaton is given as a transition graph in http://sanskrit.inria.fr/IMAGES/lexer.jpg. This development led to a generalization of the so-called *phantom phonemes* (*pretākṣara*), which are predictions on overlapping segmentations produced by the morphological generator.

Gérard Huet will give an invited lecture "From textual markup to morphological tagging to semantical annotation: the spectrum of textual annotation" at the December Delhi symposium.

On the Hyderabad side Amba Kulkarni, with collaborators Sheetal Pokar and Devanand Shukl wrote a paper "Designing a Constraint Based Parser for Sanskrit" which was accepted at the December Delhi symposium. This parser builds on similar ideas of constraint processing in the Heritage Sanskrit Parser, but uses other techniques such as a minimizer for a notion of proximity. Comparisons between the two approaches are under way.

It is hard to dissociate the Sanskrit activities of the tripartite joint team from a more international effort at Sanskrit Computational Linguistics worldwide. The Sanskrit Computational Linguistics symposium started by the joint team leaders is now well established. Its fourth occurrence is scheduled for december 2010 in Delhi. Information about the various symposia and their publication is available from http://yquem.inria.fr/~huet/EA/. Its steering committee has formed an International Sanskrit Computational Linguis-

tics Consortium (see `http://sanskrit.uohyd.ernet.in/events/2009/SCL/consortium.html`) whose aim is to mutualize joint research and resources in the field. Actually, a lot of technical debate arose from the refereeing of the numerous contributions to the Delhi Symposium. It is expected that this Symposium, with its accompanying talks and workshops, will be an important scientific event.

The joint team is organizing in December a workshop on "Sanskrit Tagging" whose purpose is to study the various proposals for morphological and syntactical markup, and to progress on the definition of standards allowing inter-operability of the various teams working on Sanskrit processing and Sanskrit philology.

In the coming collaboration for 2011, we are planning to integrate the parsers developed respectively by G. Huet and A. Kulkarni with the digital library at the Providence site. Mirror images of the various software will be established at the three sites.

2 Exchanges

Many exchanges are taking place, not just in Rocquencourt and Hyderabad, since the joint collaboration involves many other scholars. Gérard Huet visited the Department of Sanskrit Studies at Hyderabad University for two weeks in January, where he worked with Amba Kulkarni and the rest of her team, but also with Peter Scharf from the Classics Department at Brown University. He then visited TIFR in Mumbai for a few days to work with Benoît Razet and others.

Amba Kulkarni visited TIFR in February, where she presented her work on *navyanyāya*.

Vipul Mittal, an MS student at IIIT Hyderabad, worked in Spring and Summer on his MS thesis "Statistical Optimization of a Sanskrit tagger", under the joint supervision of Gérard Huet and Amba Kulkarni. He presented a poster on this work at the ACL 2010 Student Research Workshop in Uppsala in July (available as pdf at `http://www.aclweb.org/anthology/P/P10/P10-3015.pdf`). He is expected to graduate in Fall.

Benoît Razet visited Hyderabad University in August 2010 at the occasion of the International Congress of Mathematicians, where he presented his work. He worked for a week with Amba Kulkarni on the problem of incorporating quantifiers in a formal rendition of the Navyanyāya school of Indian logic.

Gérard Huet visited Hyderabad University in October for 2 weeks. He worked with Amba Kulkarni on semantic role harmony (*ākāṅkṣā-sammati*), and participated to a workshop on *kāraka* annotation with various scholars. He will visit again in December, for a week in Delhi at the occasion of the Symposium, then for two weeks in Hyderabad to work with various colleagues, then attend the South Asian Languages Analysis Roundtable (SALA) at the Central Institute of Indian Languages, Mysore, Karnataka, in January 2011.

3 Publications

The research cooperation cannot be evaluated just on joint publications, since many separate publications on similar topics by the various teams involved are actually progress reports on the respective state of the art on the various software implementations, which are cross-compared, criticized, and improved as result of this dialectical process. Furthermore the Sanskrit Computational Linguistics symposium proceedings, published as refereed volumes of the Springer Verlag Lecture Notes series, attest to the vitality and cooperative spirit of a scientific field where the joint team assumes a leading role.

The publications of Amba Kulkarni are listed at `http://sanskrit.uohyd.ernet.in/faculty/amba/`, those of Gérard Huet at `http://yquem.inria.fr/~huet/bib.html`, those of Peter Scharf at `http://www.brown.edu/Departments/Classics/people/facultypage.php?id=10044`, those of Benoît Razet at `http://www.tcs.tifr.res.in/~razet/`, Vipul Mittal's paper on a Sanskrit Segmentizer improving on Gérard Huet's sandhi analyser, under Amba Kulkarni's supervision, is available as `http://www.aclweb.org/anthology/P/P10/P10-3015.pdf`.

The series of proceedings "Sanskrit Computational Linguistics" published in the Springer Lecture Notes series is detailed at `http://yquem.inria.fr/~huet/EA/`.

The Sanskrit Library is accessible from `http://sanskritlibrary.org/`, the Sanskrit Heritage site at `http://sanskrit.inria.fr/`.