Joint team: Sanskrit Computational Linguistics
Three-year evaluation document

Creation date: 1-1-07

Consortium partners

- Centre de Paris-Rocquencourt, INRIA
- Department of Sanskrit Studies, University of Hyderabad

1 Team members on 10-1-2009

1.1 Paris-Rocquencourt Center, INRIA

Gérard Huet	Directeur de Recherche	INRIA
Benoît Razet	Étudiant en thèse	Université Denis Diderot (Paris 7)

1.2 Department of Sanskrit Studies, Hyderabad University

Amba Kulkarni	Reader & Dept Head	University of Hyderabad
Dr K Narayan Murthy Kavi	Professor, Joint faculty	University of Hyderabad
Dr K Subhramanyam	Professor, Joint faculty	University of Hyderabad
Dr Sheetal Pokar	Professor, Guest faculty	University of Hyderabad
Dr Devanand Shukla	Senior linguist	University of Hyderabad
Dr R. Anupama	Linguist	University of Hyderabad
Dr Vibhuti Nath Jha	Linguist	University of Hyderabad
Sivaja Nair	Doctoral student	University of Hyderabad
Anil Gupta	Doctoral student	University of Hyderabad
N. Shailaja	Doctoral student	University of Hyderabad

2 Cooperation history

Amba Kulkarni, a mathematician by training, did most of her research in Computational Linguistics for South Asian languages. For many years, she was working at IIT Kanpur in the team of Pr Rajiv Sangal, one of the main Indian scientists in Natural Language Understanding and Artificial Intelligence, now Head of IIIT Hyderabad. She is among other things the coordinator of morphology generation for Indian languages at the national level. She has a special interest and competence on Sanskrit linguistics and traditional grammar along the Paninian school of *Vyākaraṇa*.

Gérard Huet has had a long history of scientific cooperation with India, and when he turned his research interests towards computational linguistics around

2000, he chose the mechanical processing of Sanskrit as the choice application of the finite-state machines toolkit Zen which he was designing. This activity lead to the design and implementation of a Sanskrit Computational Linguistics platform conceived as a coordinated set of Web services around a lexical database, obtained mechanically from an original highly structured Sanskrit Heritage dictionary. In 2002 the Zen Computational Linguistics library was first released as free software. In 2003 the Sanskrit platform was first released as an interactive Web site on Internet.

A major achievement of this early system was the solution of Sanskrit segmentation, with a complete algorithm for *sandhi viccheda* implemented as a special case of rational relations over formal languages (the so-called *junction relations*). This algorithm and its proof were published for computer scientists as [11], and explained to linguists and sanskritists as [13].

By 2006 it was clear that the topic of Sanskrit Computational Linguistics was ripe for international collaboration, specially with Indian scientists and traditional scholars (*paṇḍita*). When the new department of Sanskrit Studies was created at University of Hyderabad and Amba Kulkani named as its Chair, a scientific cooperation between INRIA and University of Hyderabad was an obvious starting point for this international effort. The newly introduced INRIA scheme of "Équipe Associée" was deemed a perfect framework for the bootstrap process. The original proposal (in French) may be consulted at the URL `http://yquem.inria.fr/~huet/EA/proposition_eq_ass.html`.

In the original 3-year plan, it was proposed to organize 2 workshops on Sanskrit Computational Linguistics during the 3-year period. But soon it appeared feasible to turn the joint team technical worshops into genuine scientific events open to outside scientists, with refereed proceedings and invited lectures, complementing the technical workshop proper. Thus was born the First International Sanskrit Computational Linguistics Symposium, organized at the INRIA Paris-Rocquencourt center in October 2007. This event was sponsored by the funding of the "Équipe Associée", and complemented with additional help from both French and Indian institutions, as shown on the symposium page at `http://sanskrit.inria.fr/Symposium/`.

This event had a considerable impact, since it really started a fully international cooperation effort on this emerging new research topic. INRIA edited the Proceedings, which were integrally published in the HAL virtual library, but also edited video recordings of the talks, available at `http://sanskrit.inria.fr/Symposium/Program.html`. One of the members of the Program Committee, Pr Peter M. Scharf, from the Department of Classics of Brown University, proposed to organize the next year a second edition of this event on his NSF funding, and indeed in May 2008 the Second International Sanskrit Computational Linguistics Symposium was organized at Brown University in Providence, Rhode Island, USA. See `http://sanskritlibrary.org/Symposium/`. This event was both a very successful scientific meeting, and the occasion of an intense workshop by virtually all international scholars involved in the discipline, as witness the program at `http://sanskritlibrary.org/Symposium/Program.html`.

At this point Springer-Verlag contacted Gérard Huet, in view of regular

publishing of the refereed proceedings of the Symposium in the Lecture Notes Series. This was the turning point of the recognition of this scientific endeavour as a high-quality channel of research results in an interesting emerging field. The series was inaugurated by the joint publishing of a selection of the papers presented at the first two occurrences as a volume [15]. Soon followed by the publication of the refereed papers presented at the Third International Sanskrit Computational Linguistics Symposium [22], organized by Amba Kulkarni at Hyderabad University in January 2009 as part of our workplan `http://sanskrit.uohyd.ernet.in/Symposium/`.

This third symposium, its first occurrence within India, was a grand success because it involved the traditional scholars of Paninian grammar and allowed them to participate to the scientific discussions on an equal footing with their Western counterparts, be they linguists, philologists, or computer scientists. It is to be noted that Paninian linguistics is by no means of only historical interest. It was declared by Leonard Bloomfield, one of the leading linguists of the XXth century and the founder of modern Indo-European linguistics, that no human language had been so completely described as Sanskrit with Pāṇini's grammar (4th century BC) and the work of his successors. Thus it is of primary importance that the scholars of this still living tradition be involved in the scientific investigation of computer modelling the structures of Sanskrit. It is also to be noted that Sanskrit is not just a human language. Actually, it hardly qualifies as a natural language, since it has been strictly streamlined as a formal medium which has never been anyone's mother tongue, but is learned as a prescriptive set of rules, expounding phonology, morphology, syntax, semantics and pragmatics. It rather qualifies as a knowledge representation system, refined over centuries of use for scientific and philosophical debates, besides having a very rich litterature, both religious and profane. Thus on one hand the Sanskrit Computational Linguistics effort offers a field of investigation of far greater importance than the small number of actual locutors of the language would predict, and on the other hand the involvement of scholars trained in this tradition is of paramount importance to its success.

The conference is now well established, and its fourth occurrence is under preparation at Jawaharlal Nehru University by Pr Girish Nath Jha. It is planned for December 2010, and the call for papers has been recently issued at `http://sanskrit.jnu.ac.in/conf/4iscls/index.jsp`. The symposium has established a Steering Committee which insures its continuity, while enforcing its high scientific standards.

Regular exchange of scientists has taken place, besides the symposium/workshop meetings. Researchers from both institutions have made extended stays at the partner institution, financed by the joint team budget. For instance, Amba Kulkarni will spend one month in November 2009 at the Rocquencourt center. Gérard Huet, together with Peter Brown and a computer expert colleague, will spend one month in December 2009 at the Hyderabad site. The University of Hyderabad has determined that this collaboration was a strategic research opportunity that should be encouraged, and has proposed to INRIA to sign a Memorandom of Understanding officializing the join team collaboration within

a bilateral financing scheme for the coming 3 years.

3 Research results

At the end of the first 3-year period, a lot of progress has been accomplished in the cooperation. Each team has acquired expertise at using the tools and methods of the other site. Software has been exchanged and is regularly updated on consistent configurations. Conversion software has been written in order to align linguistic data to common representations. This permitted in particular the extensive comparison of the morphological generators by large scale benchmarks, identifying problematic derivations and leading to significant improvements in the performance of the tools on both sides. The Hyderabad site has access to considerable linguistic resources, both textual and human, which the limited human resources of the Rocquencourt side could not hope of developing independently. Conversely, the advanced parsing machinery of the INRIA Sanskrit Reader offers a parsing prototype which may integrate these linguistic resources in a comprehensive computational framework. Thus, despite the imbalance in human resources of both sides, the joint cooperation is a success.

The design of common morphological tags to insure interoperability of the various software modules has turned out to be of greater complexity than expected. Sanskrit has a very complex morphology generative scheme, which interleaves both with phonology and with syntax. Strict bottom-up computation in a layered architecture is just too restrictive, and more elaborate distributive morphology concepts have to be designed. Thus at this point we do not have yet a common representation scheme allowing the use of common databases as plugins to the software, and roundabout ways of translating them have to be endured. Furthermore, the standardization effort has been somewhat slowed down by the internationalisation of the endeavor. Thus we cannot just dictate a standard from a Rocquencourt-Hyderabad coalition, but have to elaborate a wide-consensus position, in a slower converging process.

Research achievements are judged principally by their publications. A bibliography follows, which gives the main references to the research conducted by the team, specially on the French side. The current problematics, explaining the research developments of the last 3 years and giving a set of new challenges for the next 3 years, is explained in the document (in French) which proposes the renewal of the Équipe Associée for the next 3 years, and which is available at `http://yquem.inria.fr/~huet/EA/Sanskrit09.html`. This document explains the widening of the collaboration, consistently with the growing international cooperation on the theme.

References

[1] A. Bharati, A. Kulkarni and S.S. Nair. *Use of Amarakosha and Hindi wordnet in building a Network of Sanskrit Words* In *Proceedings of ICON*

2008; 6th international Conference on NLP, 2008.

[2] P. Goyal, Amba Kulkarni, Laxmidhar Behera. *Computer Simulation of Ashtadhyayi: Some insights* In [15].

[3] Gérard Huet. From an informal textual lexicon to a well-structured lexical database: An experiment in data reverse engineering. In *Working Conference on Reverse Engineering (WCRE'2001)*, pages 127–135. IEEE, 2001.

[4] Gérard Huet. The Zen computational linguistics toolkit. Technical report, ESSLLI Course Notes, 2002.

[5] Gérard Huet. The Zen computational linguistics toolkit: Lexicon structures and morphology computations using a modular functional programming language. In *Tutorial, Language Engineering Conference LEC'2002*, 2002.

[6] Gérard Huet. Linear contexts and the sharing functor: Techniques for symbolic computation. In Fairouz Kamareddine, editor, *Thirty Five Years of Automating Mathematics*. Kluwer, 2003.

[7] Gérard Huet. Towards computational processing of Sanskrit. In *International Conference on Natural Language Processing (ICON)*, 2003.

[8] Gérard Huet. Zen and the art of symbolic computing: Light and fast applicative algorithms for computational linguistics. In *Practical Aspects of Declarative Languages (PADL) symposium*, 2003.

[9] Gérard Huet. Automata mista. In Nachum Dershowitz, editor, *Verification: Theory and Practice: Essays Dedicated to Zohar Manna on the Occasion of His 64th Birthday*, pages 359–372. Springer-Verlag LNCS vol. 2772, 2004.

[10] Gérard Huet. Design of a lexical database for Sanskrit. In *Workshop on Enhancing and Using Electronic Dictionaries, COLING 2004*. International Conference on Computational Linguistics, 2004.

[11] Gérard Huet. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *J. Functional Programming*, 15,4:573–614, 2005.

[12] Gérard Huet. Sanskrit parsing by computer. In *13th World Sanskrit Conference*, 2006.

[13] Gérard Huet. *Themes and Tasks in Old and Middle Indo-Aryan Linguistics, Eds. Bertil Tikkanen and Heinrich Hettrich*, chapter Lexicon-directed Segmentation and Tagging of Sanskrit, pages 307–325. Motilal Banarsidass, Delhi, 2006.

[14] Gérard Huet. Shallow syntax analysis in Sanskrit guided by semantic nets constraints. In *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, New York, NY, USA, 2007. ACM.

[15] Gérard Huet, Amba Kulkarni, and Peter Scharf, editors. *Sanskrit Computational Linguistics 1 & 2*. Springer-Verlag LNAI 5402, 2009.

[16] Gérard Huet and Benoît Razet. The reactive engine for modular transducers. In Kokichi Futatsugi, Jean-Pierre Jouannaud, and José Meseguer, editors, *Algebra, Meaning and Computation, Essays Dedicated to Joseph A. Goguen on the Occasion of His 65th Birthday*, pages 355–374. Springer-Verlag LNCS vol. 4060, 2006.

[17] Gérard Huet and Benoît Razet. Computing with relational machines. Tutorial, ICON 2008, Pune.

[18] A. Kulkarni. *Grammarians' interface for English Parsers*. National seminar on 'Sanskrit for Innovations', Center for Advanced Studies in Sanskrit, University of Pune, Pune, 2006.

[19] A. Kulkarni. *Human Understandable Machine Learning*. National seminar on 'Sanskrit for Innovations', Keynote Speech in NLPAI: National Workshop on AI, NCST, Mumbai, 2006.

[20] A. Kulkarni and Sriram Chaudhury. *English Parsers: Some Information based observations* Workshop on MRCS-07, at 20th IJCAI, Hyderabad, 2007.

[21] A. Kulkarni. *'Subject' in English is abhihita*. 14th Sanskrit World Conference, Kyoto, Japan, 2009.

[22] Amba Kulkarni and Gérard Huet, editors. *Sanskrit Computational Linguistics 3*. Springer-Verlag LNAI 5406, 2009.

[23] Anil Kumar, V. Sheebasudheer and A. Kulkarni. *Sanskrit Compound Paraphrase Generator*. In *Proceedings of ICON 2009, 7th International Conference on NLP*, forthcoming December 2009.

[24] A. Kulkarni, R. YousufZai, and Parvez Ahmad. *Urdu-Hindi-Urdu Machine Translation: Some problems* In *Proceedings of Conference on Language and Technology*, Lahore, Pakistan, 2009.

[25] S. S. Nair, P. Swain and A. Kulkarni. *Developing network of Sanskrit words across Part-Of-Speech categories*. In *Proceedings of National Seminar on Computer Science and its Applications in Traditional Shastras*, Rashtriya Sanskrit Vidyapeeth, Tirupati; to appear, 2009.

[26] Benoît Razet. Finite Eilenberg machines. In O.H. Ibarra and B. Ravikumar, editors, *Proceedings of CIIA 2008*, pages 242–251. Springer-Verlag LNCS vol. 5148, 2008. `http://gallium.inria.fr/~razet/fem.pdf`

[27] Benoît Razet. Simulating finite Eilenberg machines with a reactive engine. In *Proceedings of MSFP 2008*. Electric Notes in Theoretical Computer Science, 2008. `http://gallium.inria.fr/~razet/PDF/razet_msfp08.pdf`

[28] Benoît Razet. *Effective Eilenberg Machines*. PhD thesis, University Denis Diderot (Paris 7), forthcoming November 2009.

[29] P. Shukl, D. Shukl and A. Kulkarni. *Vibhakti Level divergences between Sanskrit and Hindi*. In *Proceedings of AICL, 31st All India Conference on Linguistics*, forthcoming December 2009.