Volume-II, No. 1, Feb -2015, pp. 44-49

IL-Indian Languages Machine Translation Systems in India and Role of Sanskrit in Translation among Indian Languages (IL)

Mr. Uday Narayan Singh,
Dr. N Shankaranarayana Sastry

Introduction

The present paper deals with the role of Sanskrit in translation among Indian languages and Indian Language (IL) machine translation systems in India. Sanskrit enjoys a place of pride among Indian languages in terms of technology solutions that are available for it within India and abroad. Also in terms of committed groups working on in a mission mode all over the globe. The Indian government through its various agencies has been heavily funding other Indian languages for technology development but the funding for Sanskrit has been slow for a variety of reasons. Despite that, the work in the field has not suffered. The following sections do a survey of the language technology R&D in Sanskrit and other Indian languages.

1.1 Indian Languages MT Systems in India

Centre for Development of Advanced Computing (C-DAC) with funding from Ministry of Communications and Information Technology, Government of India. C-DAC is committed to design, develop and deliver advanced computing solutions for human advancement. The Applied Artificial Intelligence (AAI) group at C-DAC is working on some of the fundamental applications in the field of Natural Language Processing, Machine Translation, Intelligent Language teaching and Decision Support Systems.[1]

In 1990-91, Government of India launched TDIL (Technology Development for Indian Language) program for development *of corpora*, OCR, *Text-to-Speech*, Machine Translation and language processing tools. India is a multi-lingual country with 22 constituent languages with 10 different scripts. Only 5% of Indian population can work in English. Two percent of the world's languages are becoming extinct every year. Four European languages (English, German, French and Spanish) comprise more than 80 percent of all book translations. There is a worldwide unquantifiable erosion of cultural participation and innovation. With the loss of a language, we lose art and ideas, scientific information and innovative capacity, knowledge about medical plants and preparations that could cure maladies. Gap between scientific contributions in linguistic communities is widening. Every year, about 46,000 journals and over 80,000 books in science & technology are published. Most of this is in English, and negligible in the languages of developing economies [INSDOC, 2000] According to Dr. Om Vikas[2], the Development of Language Technology in India may be categorized in three phases as follows-

- **Technology Phase (1976-1990)** : Focus was on Adaptation Technologies; abstraction of requisite technological, designs and competence building in R&D institutions.
- **Technology Phase (1991-2000) :** Focus was on developing Basic Technologies generic information processing tools, interface technologies and cross-compatibility conversion
- **Technology Phase (2001-2010) :** Focus is on developing Creative Technologies in the context of convergence of computing, communication and content technologies. Collaborative technology development is being encouraged

The **Anusaaraka**[3] project originated at IIT Kanpur and later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL. Today, the Language Technology Research Centre (LTRC) at IIIT Hyderabad is attempting an English-Hindi Anusaaraka MT system.

Anusaaraka is using principles of Paninian Grammar (PG) and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and target languages. The project has developed Language Accessors from Punjabi, Bengali, Telugu, Kannada and Marathi into Hindi. The

approach and lexicon is general, but the system has mainly been applied for children's stories.[4]

The Anusaaraka does not do sentential analysis because it requires a large amount of data to be prepared. Also, since the Indian languages are close, the 80-20 rule applies to *vibhakti.* Use of *vibhakli* produces 80 percent 'correct' output with only 20 percent effort.[5] Sentential parser can be incorporated when large lexical databases are ready. The next stage of processing is that of the mapping block. This stage uses a noun *vibhakti* dictionary. For each word group, the system finds a suitable root and *vibhakti* in the target language. Thus, it generates a local word group in the target language.

The local word groups in the target language are passed on to a local word splitter (LWS) followed by a morphological synthesiser (GEN). The LWS splits the local word groups into elements consisting of root and features. Finally, the GEN generates the words from a root and the associated grammatical features.[6]

In spite of difficulties in MT, the Anusaaraka is believed to be useful to overcome the language barrier in India today.[7] Anusaaraka systems among Indian languages are designed by noting the following two key features:

1. In the Anasaaraka approach, the load between the reader and the machine is divided in such a way that the aspects, which are difficult for the reader, are handled by the machine and aspects, which are easy for the reader, are left to him. Specifically, reader would have difficulty learning the vocabulary of the language, while he would be good at using general background knowledge needed to interpret any text.
2. Among Indian languages, which share vocabulary, grammar, pragmatics, etc. the task is easier. For example, in general the words in a language are ambiguous, but if the languages are close to each other, one is likely to find a one to one correspondence between words where the meaning is carried across from source language to target language.

In the Anusaaraka approach, the reader is given an image of the source text in the target language by faithfully representing whatever is actually contained in the source language text. So, the task boils down to presenting the information to the user in an appropriate form.

Some amount of training will be needed on the part of the reader to read and understand the output. This training will include teaching of notation, some salient features of the source language and is likely to be about 10 percent of the time needed to learn a new language. For example, among Indian languages it could be of a few weeks duration, depending on the proficiency desired. It could also occur informally as the reader uses the system and reads its output, so the formal training could be small.[8]

Shakti[9] system for M (A)T from English to three Indian languages (Hindi, Marathi and Telugu) is developed by LTRC, IIIT Hyderabad. Shakti machine translation system has been designed to produce machine translation systems for new languages rapidly. It has been already developed for English to three different Indian languages—Hindi, Marathi and Telugu. The limited release of 'Shakti-kit' was done in ICON-2003[10]. The system is so designed that many of the benefits of improvement to the system flow automatically to outputs in all the languages. The Shakti is also designed to take ready made sub-systems either as black boxes or as open source software and incorporate them into it self. The simplicity of the overall architecture makes it easy to do so. Available English analysis packages have been extensively adapted by the Shakti. A number of system organisation principles have been used, which have led to the rapid development of the system. While the principles by themselves might not appear to be new, their application to MT in this manner is unique.

MaTra is an ongoing project at C-DAC, Mumbai, and has been funded by TDIL. It aims at machine-assisted translation from English into Hindi, essentially based on a transfer approach using a frame-like structured representation. The focus is on the innovative use of man-machine - synergy - the user can visually inspect the analysis of the system and provide disambiguation information using an intuitive GUI, allowing the system to produce a single correct translation.[11] The system uses rule-bases and heuristics to resolve ambiguities to the extent possible - for example, a rule-base is used to map English prepositions into Hindi postpositions. The system can work in a fully automatic mode and produce rough translations for end users, but is primarily meant for

translators, editors and content providers. Currently, it works for simple sentences and work is on to extend the coverage to complex sentences. The MaTra lexicon and approach is general-purpose, but the system has been applied mainly in the domains of news, annual reports and technical phrases.

The MaTra is available in two versions: MaTra Pro - the Professional Translator's Tool; and MaTra Lite - automatic On-line Translator. MaTra Pro is more appropriate for serious translation - it is a Translator's Tool and allows the user to assist the system in generating more accurate translations. It uses a simple, intuitive GUI for interaction and allows customization of the lexicon to specific application domains. The aim is for the machine to try and translate the simpler or more routine texts, and free the human translator to focus on the more difficult and creative tasks. MaTra Pro is currently available under non-exclusive license. Features:

- Auto, Semi-Auto and Manual Modes;
- Intuitive GUI for disambiguation
- User-customizable lexicon
- More Accurate Translation

It is ideal for:

- Editors, content-providers
- Professional translators

MaTra Lite expects no interaction from the user and is, therefore, able to do only rough translation. The entire range of complexity of English inputs is not handled. Also, due to some of the choices taken independently by the system, the translations may not be entirely correct. A version of MaTra Lite is being offered as a free, experimental Web-based service hosted at CDAC Mumbai's (formerly NCST) website (http://www.ncst.ernet.in/matra/) and accessible through a browser. The main features are:

- Simple web-based interface
- Automatic, end-to-end operation
- General, approximate translation

And it ideal for:

- Non-Hindi speakers learning Hindi
- Web users unfamiliar with English

The prototype for assertive sentences with one verb group is now being extended to handle compound-complex sentences.[12]

The **Mantra** project has been developed by C-DAC, Bangalore. The project has been funded by TDIL and later by the Department of Official Languages. Mantra becomes part of Smithsonian Institution's National Museum of American History.

The project is based on the TAG formalism from University of Pennsylvania, USA. A sub-language English-Hindi MT system has been developed for the domain of gazette notifications pertaining to government appointments. In addition to translating the content, the system can also preserve the formatting of input Word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain. Recently, work has been initiated on other language pairs such as Hindi-English and Hindi-Bengali, as well as on extending to the domain of parliament proceeding summaries.

Anuvaadak[13] 5.0 Machine Translation system from English to Hindi is developing by Super Info soft Pvt. Ltd; this software is compatible with Microsoft Windows 95/98, Windows NT. Some important features of this system is as filling and Printing facility, inbuilt grammar checker for both language, Includes 100 Hindi free fonts, User-friendly with pull-down menus, User can add more multiple Hindi meanings to the existing ones also.

Oriya Machine Translation System (OMT) translating from English to Oriya is being developed by the Research Centre of the Department of Computer Science, Utkal University, and Vanivihar14. The architecture of the OMT is divided into six parts: Parser, Translator, OMT System, OMT Database, Disambiguator and the Software tools. The heart of the system is the OMT database (bilingual dictionary). In this database, various information is stored: English word, category, tense,- Oriya meaning of the word, etc. The Oriya meaning of the words has been stored in simple ISC1I format. The system is developed using Java programming languages and MySQL as the database.

The parser takes the English sentence as input from the OMT system and then every sentence is parsed according to various rules of parsing with the help of OMT database. During parsing, in some cases it is taking the help of morphological analyser. The translator fakes the parsed structure of sentences from the OMT system as input and then it performs the task of translation with the

help of OMT database. This translator part is taking the help of tense analysis and verb tagging module. The disambiguator module is doing the task of disambiguation with the help of frequencies obtained from Corpora (Oriya Corpora) by using the n-gram module.

Tamil University Machine Translation System[15] (TUMTS) was one of the main projects of Tamil University, Tanjore, during 1980s. As a beginning a machine oriented translation, involving Russian-Tamil was initiated during 1983-1984 under the leadership of the Vice-Chancellor Dr. V.I Subramaniam. It was taken up as an experimental project to study and compare Tamil with Russian in order to translate Russian scientific text into Tamil. Hence, the goal was kept minimal and the scientific text belonging to a specific domain was used as SL input. A team consisting of a linguist, a Russian language scholar and a computer scientist was formed to work on this project. During the preliminary survey, both Russian SL and Tamil were compared thoroughly for their style, syntax, morphological level etc.

Bharathidasan University, Tamilnadu is working on **Tamil- Malayalam Machine Translation[16] System.** Languages belonging to the same family will naturally have similar morphological and syntactical structure. The cultural aspects of the speakers of both the languages will be either similar or easy for the translator to understand. With this hypothesis, the translation of Tamil-Malayalam translation they have started. For this MT they have developing four components as follows-

Lexical database- This will be a bilingual dictionary of root words. All the noun roots and verb roots are collected.

Suffix database- Inflectional suffixes, derivative suffixes, plural markers, tense markers, sariyai, case suffixes, relative participle markers, verbal participle markers etc will be compiled.

Morphological Analyzer- It is designed to analyze the constituents of the words. It will help to segment the words into stems, inflectional markers.

Syntactic Analyzer- The syntactic analyzer will find the syntactic category like Verbal Phrase, Noun Phrase, Participle Phrase etc. This will analyze the sentences in the source text.

The Anna University K.B. Chandrasekhar Research Centre at Chennai was established in January 2000 and is active in the area of Tamil NLP. A **Tamil-Hindi** language accessor has been built using the Anusaaraka formalism described above. Recently, the group **has** begun work on an **English-Tamil MT system.**

The Natural Language Processing (NLP) group focuses on developing Tools, Technologies, Products and Systems to facilitate the use of computers and the Internet for day-to-day life. The group also works on building lexical resources such as dictionaries, Word Net and tagged corpora that are essential for researchers working on various areas of NLP [17]. Both linguists and computer scientists are members of the NLP group. The Tamil-Hindi MAT has the following structure? [17]

- Morphological analyzer of source language
- Mapping unit
- The target language generator

Major Machine Translation Projects in India[18]

Project Name	Languages	Domain/ Main Application	Approach/ Formalism	Strategy
Anglabharati (IIT-K and C-DAC, N)	Eng-II (Hindi)	General (Health)	Transfer/Rules (Pseudointerlingua)	Post-Edit
Anusaaraka (IIT-K and University of Hyderabad)	IL-IL (5IL> Hindi) [5IL; Bengali, Kannada, Marathi, Punjabi, and Telugu)	General (Children)	LWG mapping /PG	Post-edit
MaTra (C-DAC, M)	Eng-IL (Hindi)	General (News)	Transfer/ Frames	Pre-edit
Mantra (C-DAC, B)	Eng-IL (Hindi)	Government Notifications	Transfer/XTAG	Post-edit
UCSG MAT (University of Hyderabad)	Eng-IL (Kannada)	Government Circulars	Transfer/UCSG	Post-edit
UNL-MT (IIT-B)	Eng, Hindi, Marathi	General	Interlingua/UNL	Post-edit

Tamil Anusaaraka (AU-KBC, C)	IL-IL (Tamil-Hindi)	General (Children)	LWG mapping/ PG	Post-edit
MAT (Jadavpur University)	Eng-IL (Hindi)	News Sentences	Transfer/Rules	Post-edit
Anuvaadak (Super Infosoft)	Eng-IL (Hindi)	General	[Not Available]	Post-edit
StatMT (IBM)	Eng-IL	General	Statistical	Post-edit

Major approaches to MT

The following table summarizes the advantages/disadvantages[19] of each MT approach-

Approaches	Advantages	Disadvantages
Rule-Based	1. easy to build an initial system 2. based on linguistic theories 3. effective for core phenomena	1. Rules are formulated by experts. 2. difficult to maintain and extend 3. ineffective for marginal phenomena
Knowledge-Based	1. based on taxonomy of knowledge 2. contains an inference engine 3. inter lingual Representation	1. hard to build knowledge hierarchy 2. hard to define granularity of knowledge 3. hard to represent knowledge
Example Based	1. extracts knowledge from corpus 2. based on Translation Patterns in corpus 3. reduces the human cost	1. similarity measure is sensitive to system 2. search cost is expensive 3. knowledge acquisition is still problematic
Statistics-Based	1. numerical knowledge 2. extracts knowledge from Corus 3. reduces the human cost 4. model is mathematically grounded	1. no Linguistic background 2. Search cost is expensive 3. hard to capture long distance phenomena

1.2 Role of Sanskrit in Translation among Indian Languages (IL)

Sanskrit is probably the oldest and genealogically most connected language of the Indian sub-continent. Besides, almost all major Indian languages have inherited lexical. Linguistic. And stylistic features from Sanskrit. The common cultural heritage of the speakers of Indian languages also makes Sanskrit a connecting link between them. Linguists like Emenou have explored the possibility of a 'linguistic area' in India. The fact that there are obvious linguistic similarities among Indian languages, need to be exploited for machine translation among Indian languages. There is an urgent need to develop linguistic resources and tools based on Paninian frame-work using Sanskrit as Interlingua for MT among Indian languages.

Sanskrit has been a 'donor language' in the Indian context. Not only have the modern Indo-Aryan languages liked Hindi. Panjabi, Bangla, Marathi, Gujrati (These languages are called *janya-bhdsas*, the languages evolved from Sanskrit) etc. but also Dravidian languages like Kannada, Telugu, Malayalam and Tamil (to some extent) are beneficiaries of vast vocabulary of Sanskrit. The structure and semantics of these languages owe a great deal to Sanskrit. Whenever there is a need for a new technical word to be coined that can be accepted and integrated in these languages, Sanskrit is the sole source. The Sanskrit words appearing in these languages without any change are called--*tatsama* (equivalent to itself) words. It is said that about 36% of the words in Bangla are *tatsamas.*[20] Sanskrit words which are integrated into these languages with some modifications are called - *tadbhava-s* (derived from it). For e.g., *rathya- rasta* (Hindi); *prakasa* (Sanskrit) *parkas* (Punjabi); *maharastra - maratha* (Marathi); *laksana* (Sanskrit) - *ilakkanam* (Tamil); *pustaka* (Sanskrit) - *hottige* (Kannada); *vithi* (Sanskrit)- *vidi* (Telugu).

The traditional grammars of the Modern Indian languages which are used till date are based on the Paninian structure. The grammatical categories in these languages are classified in Paninian way and bear the same names. The nominal paradigms are treated in seven cases as in Sanskrit. The south

Indian languages though belong to a different family of Dravidian languages they have been highly influenced by Sanskrit

Conclusion

The conclusion derived from various studies on Indian Languages MT systems in Indian. In the field of NLP, Paninian frame-work has been applied on most of the Indian languages, and is tested to be best for the Indian languages.[21] There have been attempts to use Sanskrit as Interlingua for MT among Indian languages.[22] Such being the close affinity between Sanskrit and other Indian languages there is a need to speed up research on applying the Sanskrit *Sastraic* techniques of Paininian Grammar, Navya Nyaya and Mimamsa for language processing in Indian languages especially in automatic translation systems among Indian languages.

References

1. http://www.cdacindia.com/htm/about/success/mantra.asp (accessed on 14.03.05)
2. Dr. Om Vikas UNESCO Meeting on Multilingualism for Cultural Diversity and Universal Access in Cyberspace : An Asian perspective, 6-7 May 2005.
3. Anusaaraka, http://www.iiit.net/ltre/Anusaaraka/anu home.html (accessed: 7 April 2005).
4. Bharati, Akshar; Kulkarni, Amba P; Chaitanya, Vineet: Sangal, Rajeev and Rao. G. Umamaheshwara. 2000, *Anusaaraka: Overcoming the Language Barrier in India. In Anuvaad.* Sage Publishers, New Delhi.
5. Bharati, Akshar; Chaitanya, Vineet and Sangal, Rajeev, 1999, *Natural Language Processing. A Paninian Prospective.* Prentice Hall of India.
6. Bharati, Akshar; Chaitanya, Vineet and Sangal, Rajeev, 1999, *Natural Language Processing. A Paninian Prospective.* Prentice Hall of India.
7. Bharati, Akshar; Kulkarni, Amba P; Chaitanya, Vineet: Sangal, Rajeev and Rao. G. Umamaheshwara, 2000, *'Anusaaraka: Overcomng the Language Barrier in India. In Anuvaad".* Sage Publishers, New Delhi.
8. Bharati, Akshar, Kulkarni, Amba P; Chaitanya, Vineet: Sangal, Rajeev and Rao, G. Umamaheshwara. 2000.' *anusaaraka: Overcoming the Language Barrier in India, In Anuvaad".* Sage Publishers, New Delhi.
9. Rajeev Sangal, http://www.elitexindia.com/paper2004/rajeevsangal .pdf (accessed: 7 April 2005).
10. http://ltrc.iiit.net/showfile.php?filename=projects/shakti.php (accessed on 16.03.05)
11. http://www.ncst.ernet.in/matra/about.shtml (accessed on 16.03.05);
12. Mehta, Vivek and Rao, Durgesh, 2001, '*Natural Language Generation of Compound Complex Sentences for English-Hindi Machine-Aided Translation'*, National Centre for Software Technology, Mumbai.
13. Anuvaadak. http:tdil.mit.gov.in/TDIL Jan-April 2004/super%20 infosof t%20pvt%20ltd.pdf (accessed : 7 April 2005).
14. Mohanty, S. and Balabantaray, R.C. 2004. *Machine Translation System (Orya)',* Processing of Symposium on Indian Morphology, Phonology and Language Engineering SIMPLE 2004, held on IIT Karagpur, March, 2004.
15. Dr. K.C. Chellamuthu. http://www.infitt.org/ti2002/papers/16 CHELLA.PDF (accessed on 7th April 2007).
16. Dr. Radha Chellappan, An Approach to Tamil- Malayalam Machine Tanslation, site: http:www.infitt.org/ti2003/papers/03 rchellap.pdf (accessed on 7th April 2007).
17. Shanmugam, B Kumara, 2002, *'Machine Translation as Related to Tamil',* Tamil Internet 2002, INFITT, San Francisco, CA, USA, September, 2002;
18. Salil Badodekar, "Translation Resources, Services and Tools for Indian Languages".
19. Chen, Kuang-Hua & Hsin-Hsi Chen. 1996. A Hybrid Approach to Machine Translation System Design, Computational Linguistics and Chinese Language Processing, Vol I no 1. August, 1996, pp. 159-182.
20. Akkas. Abu Jar M., 'Collocation of Pahela and Pratham'. Editorial, The Holiday (International Edition) http://www.weeklyholdiay.net/ 111002/edit.html (accessed on : 07-04-2007).
21. Bharati, Akshar, Chaitanya, Vineet & Sangal, Rajeev, 1999. Natural Language Processing: A Paninian Perspective. Printice-Hall of India Pvt Ltd., New Delhi- 01 p. xiii.
22. Sinha, R.M.K. & Jain, A., 'Angla-Hindi: An English to Hindi Machine-Aided Translation System', MT Summit IX, New Orleans, USA. http://www.amtaweb.org/summit/MTSummit/Final papers/36-sinha-final.pdf (accessed on: 07-04-2007).