Translation Resources, Services and Tools for Indian Languages

Salil Badodekar ***salil@cse.iitb.ac.in***

Computer Science and Engineering Department
Indian Institute of Technology, Mumbai, 400019, India.

Contents

List of Tables

Abstract
This paper surveys translation resources, services and tools available in the 18 officially recognized Indian languages. Major machine translation projects, language corpora and available resources, services and tools are covered. The resources include concordance, corpora (with and without annotation), dictionary, lexicon, thesaurus, and WordNet. The tools include chunker, language accessor, morphological analyser, parser, semantics analyser, syntax analyser, speaker verification system, speech recognition system, speech synthesizer, spell-checker, tagger, text to speech synthesiser, word processor, and word-sense disambiguator. The services are tools available online.

Keywords
language corpora, language resources, language technology, machine translation, translation tools.

Motivation

India has 18 officially recognized languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali**,** Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. Clearly, India owns the language diversity problem. In the age of Internet, the multiplicity of languages makes it even more necessary to have sophisticated machine translation systems. Many machine translation projects and related activities are going on in the country and abroad. Hence, it is essential to find out the current state of technology.

Scope

This document includes information about the following:

- Resources: concordance, corpora (with and without annotation), dictionary, lexicon, thesaurus, WordNet.

- Services: these are tools available online.

- Tools: chunker, language accessor, morphological analyser, parser, semantics analyser, syntax analyser, speaker verification system, speech recognition system, speech synthesizer, spell-checker, tagger, text to speech synthesiser, word processor, word-sense disambiguator.

This document does not include information about the following: bulletin board system, CD authoring tool for Indian language documents, font, font-related issue, keyboard driver, multilingual e-mail client, search engine, standard, web based e-mail service.

Major Machine Translation Projects in India

A table from [6] is reproduced here with some updations.
Please see the disclaimer near the end of the document.

Project Name	Languages	Domain/ Main Application	Approach/ Formalism	Strategy	Brief Description
Anglabharati (IIT-K and C-DAC, N)	Eng-IL (Hindi)	General (Health)	Transfer/Rules (Pseudo-interlingua)	Post-edit	1
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	2
MaTra (C-DAC, M)	Eng-IL (Hindi)	General (News)	Transfer/Frames	Pre-edit	3
Mantra (C-DAC, B)	Eng-IL (Hindi)	Government Notifications	Transfer/XTAG	Post-edit	4
UCSG MAT (University of Hyderabad)	Eng-IL (Kannada)	Government circulars	Transfer/UCSG	Post-edit	5
UNL MT (IIT-B)	Eng, Hindi, Marathi	General	Interlingua/UNL	Post-edit	6
Tamil Anusaaraka (AU-KBC, C)	IL-IL (Tamil-Hindi)	General (Children)	LWG mapping/PG	Post-edit	7
MAT (Jadavpur University)	Eng-IL (Hindi)	News Sentences	Transfer/Rules	Post-edit	8
Anuvaadak (Super Infosoft)	Eng-IL (Hindi)	General	[Not Available]	Post-edit	9
StatMT (IBM)	Eng-IL	General	Statistical	Post-edit	10

ASR, M — Academy of Sanskrit Research, Melkote
AU-KBC, C — Anna University's K. B. Chandrasekhar Research Centre, Chennai
C-DAC, B — Centre for Development of Advanced Computing, Bangalore
C-DAC, M — Centre for Development of Advanced Computing, Mumbai (Erstwhile NCST)
C-DAC, N — Centre for Development of Advanced Computing, Noida (Erstwhile Electronics, Research and Development Centre of India)
C-DAC, MH — Centre for Development of Advanced Computing, Mohali (Erstwhile CEDTI)
C-DAC, T — Centre for Development of Advanced Computing, Thiruvananthapuram (Erstwhile Electronics, Research and Development Centre of India)

CEERI, D	Central Electronics Engineering Research Institute, Delhi
CIIL, M	Central Institute for Indian Languages, Mysore
IBM	International Business Machines, U.S.A.
IIT-B	Indian Institute of Technology, Mumbai
IIT-K	Indian Institute of Technology, Kanpur
ISI-K	Indian Statistical Institute, Kolkata
JNU, ND	Jawahar Lal Nehru University, New Delhi
LTRC, IIIT, H	Language Technologies Research Center, IIIT, Hyderabad
TDIL	Technology Development for Indian Languages

Contact Information about the Major Machine Translation Projects in India

Information given in [6] proved useful in creating the table below.
Please see the disclaimer near the end of the document.

Project and Agency	URL	Contact Person(s)	Email
Anglabharati (IIT-K, C-DAC, NOIDA)	http://www.cse.iitk.ac.in/users/langtech/anglabharti.htm	Prof. R. M. K. Sinha	<rmk@cse.iitk.ac.in>
Anusaaraka (IIT-K, University of Hyderabad)	http://www.iiit.net/ltrc/Anusaaraka/anu_home.html	Prof. Rajeev Sangal Prof. G. U. Rao	<sangal@iiit.net> <guraosh@uohyd.ernet.in>
MaTra (C-DAC, M)	http://www.ncst.ernet.in/matra/ http://www.ncst.ernet.in/matra/about.shtml	Durgesh Rao MaTra Team	<durgesh@ncst.ernet.in> <matra@ncst.ernet.in>
Mantra (C-DAC, B)	http://www.cdacindia.com/html/about/success/mantra.asp	Dr. Hemant Darbari	<darbari@cdac.ernet.in>
UCSG MAT (University of Hyderabad)	http://www.uohyd.ernet.in/	Prof. K Narayana Murthy	<knmcs@uohyd.ernet.in>
UNL MT (IIT-B)	http://www.cfilt.iitb.ac.in/	Prof. Pushpak Bhattacharyya	<pb@cse.iitb.ac.in>
Tamil Anusaaraka (AU-KBC, C)	http://www.au-kbc.org/frameresearch.html	Prof. C. N. Krishnan	<cnkrish@au-kbc.org>
MAT (Jadavpur University)	http://www.jadavpur.edu/	Prof. Sivaji Bandyopadhyay	<ilidju@cal2.vsnl.net.in>
Anuvaadak (Super Infosoft)	http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak	Ms. Anjali Rowchowdhury	<anjalir@del16.vsnl.net.in>
StatMT (IBM)	http://www.research.ibm.com/irl/projects/translation.html	Not Available	Not Available

List of Resource Centres

A table from http://tdil.mit.gov.in/resource_centre.htm is reproduced with some changes.

Language(s)	Resource Centre Associated With
Assamese, Manipuri	Indian Institute of Technology, Guwahati
Bengali	Indian Statistical Institute, Kolkata
Foreign Languages (Japanese, Chinese) & Sanskrit (Language Learning Systems)	Jawaharlal Nehru University, New Delhi
Gujarati	MS University, Baroda
Hindi, Nepali	Indian Institute of Technology, Kanpur
Kannada, Sanskrit (Cognitive Models)	Indian Institute of Science, Bangalore
Malayalam	C-DAC, Thiruvananthapuram
Marathi, Konkani	Indian Institute of Technology, Mumbai
Oriya	Utkal University, Department of Computer Science and Application
Punjabi	Thapar Institute of Engg. & Tech., Patiala
Tamil	Anna University, Chennai
Telugu	University of Hyderabad, Hyderabad
Urdu, Sindhi, Kashmiri	CDAC, Pune

Development of Language Corpora in Indian Languages

A table from [2] is reproduced here with some changes.

A corpus of size 3 million words was built for each language mentioned in the following table. The duration for first five projects was 1991-1995, and for the last one, it was 1992-1995.

Please see the disclaimer near the end of the document.

Language(s)	Implementing Agency
Hindi, English, Punjabi	Indian Institute of Technology, New Delhi
Kannada, Malayalam, Tamil, Telugu	Central Institute of Indian Languages, Mysore, Karnataka
Marathi, Gujarati	Deccan College, Pune, Maharashtra
Oriya, Assamese, Bangla	Indian Institute of Applied Language Sciences, Bhubaneswar, Orissa
Sanskrit	Sampurnananda Sanskrit University, Varanasi, Uttar Pradesh
Urdu, Sindhi, Kashmiri	Aligarh Muslim University, Aligarh, Uttar Pradesh

Available Resources, Services and Tools

Legend
- Not applicable or not mentioned
15 IL 15 (of the 18) officially recognized Indian languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sanskrit, Tamil, Telugu, Urdu.
PoA A 'Y' in this column indicates that the tool is a Part of an Application. A tool may be developed as a part of an application and as such, may not be available separately.

Please see the disclaimer near the end of the document.

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Chunker	Y	Shakti	-	LTRC, IIIT, H	-	-
Concordance	Y	Sanskrit Authoring System	Sanskrit	C-DAC, B	Coming up	Windows
Corpora (Annotated)	N	-	Hindi, Telugu	LTRC, IIIT, H	Coming up	-
Corpora (tagged with grammatical category)	N	-	Hindi	LTRC, IIIT, H	http://www.tdil.mit.gov.in/download/menu.htm	MS-DOS version 6.0 or higher File Size 131 MB approx.
Corpora	N	-	15 IL	TDIL	-	-
Language accessor	N	Anusaaraka	English to Hindi	LTRC, IIIT, H	-	-
Lexicon	N	WordNet	Bengali	IIT Kharagpur	http://www.mla.iitkgp.ernet.in/technology.html#ilt (info. only)	-
Lexicon	N	WordNet	Hindi	IIT, Mumbai	http://www.cfilt.iitb.ac.in/wordnet/webhwn/	-
Lexicon	N	WordNet	Himachali, Kannada, Kashmiri, Punjabi, Urdu	CIIL, Mysore	Coming up	-

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Lexicon	N	WordNet	Konkani	IIT, Mumbai	Coming up	-
Lexicon	N	WordNet	Marathi	IIT, Mumbai	http://www.cfilt.iitb.ac.in/wordnet/webmwn/	-
Lexicon	N	WordNet	Sanskrit	Utkal University	http://www.ilts-utkal.org/orinet.htm	-
Lexicon	N	WordNet	Tamil	AU-KBC, C	http://www.au-kbc.org/research_areas/nlp/projects/tamil_wordnet.html (info. only)	-
Lexicon	N	WordNet	Telugu	University of Hyderabad	-	-
Lexicon / Dictionary	N	Trilingual Dictionary	English, Hindi and Malayalam	C-DAC, T	http://www.malayalamresourcecentre.org/Mrc/products/trilingual.html	-
Lexicon / Dictionary	N	Bilingual Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	Learners Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	Recall Voice Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	Pictorial Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	Phonetic Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	Etymological Dictionary	-	CIIL	http://www.ciil.org/development/	-

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Lexicon / Dictionary	N	General Purpose Dictionary	-	CIIL	http://www.ciil.org/development/	-
Lexicon / Dictionary	N	-	Bengali-English	Indian Statistical Institute, Kolkata	http://www.isical.ac.in/~rc_bangla/products.html (info. only)	-
Lexicon / Dictionary	N	Winki	English-Hindi	S.R.G. Systems Pvt. Ltd., Software Research Group	-	-
Lexicon / Dictionary	Y	Bilingual dictionary	Oriya <=> English	Utkal University	http://www.ilts-utkal.org/e_dic.htm	Windows-98/2000/NT, Linux
Lexicon / Dictionary	Y	Amarakosha	Sanskrit	C-DAC, B	Coming up	-
Lexicon / Dictionary (Databases: 690 Avyayas, 26,000 Nominal stems, 600 Verbal roots, krdanata forms of 600 verbal roots, 5 Taddhita suffixes)	Y	-	Sanskrit	ASR, M	-	DOS platform with GIST card; and is being ported to Windows
Lexicon: E-dictionary	N	Shabdaanjali	English to Hindi	LTRC, IIIT, H	http://www.iiit.net/ltrc/Dictionaries/Dict_Frame.html (download)	-

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Lexicon: E-dictionary	Y	Sanskrit Authoring System	Sanskrit	C-DAC, B	Coming up http://www.cdacindia.com/html/connect/3q2000/art10a.htm (product info. only)	Windows
Lexicon: Transfer Lexicon and Grammar	N	TransLexGram	English to Hindi	LTRC, IIIT, H	Coming up	-
Machine translation service	N	Shakti	English to Hindi	LTRC, IIIT, H	http://216.236.98.137/~shakti/	-
Machine translation service	N	-	English to UNL (interlingua)	IITB	http://laiir.cse.iitb.ac.in/eng_unl_anal.html	-
Machine translation service	N	-	Hindi to UNL (interlingua)	IITB	http://www.cfilt.iitb.ac.in/eng-hin-mt/	-
Machine translation service	N	Anglabharati	English to Hindi	IIT-K, C-DAC, NOIDA	http://anglahindi.iitk.ac.in/index2.html http://anglahindi.iitk.ac.in/newpages/footer.htm	-
Machine translation software	N	Anuvaadak	English to Hindi	Super Infosoft	http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak (product info. only)	-

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Machine translation software	N	Oriya Machine Translation System (OMTrans)	English to Oriya	Utkal University, Vanivihar	http://www.ilts-utkal.org/omt.htm (product info. only)	-
Morphological analyser	Y	-	Sanskrit	C-DAC, B	Coming up	-
Morphological analyser	N	-	Hindi, Kannada, Marathi, Punjabi, and Telugu	IIT-K, University of Hyderabad	Coming up	-
Morphological analyser	N	-	Hindi, Kannada, Marathi, Punjabi, and Telugu	LTRC, IIIT, H	http://www.iiit.net/ltrc/morph/index.htm (download)	Linux, Perl, GDBM, Flex, Perl enabled vim (only for Telugu)
Morphological analyser	Y	-	Sanskrit	C-DAC, B	Coming up	-
Morphological analysis and generator	Y	-	Sanskrit	ASR, M	-	DOS with GIST card; and is being ported to Windows
Morphological learner	N	-	Any	LTRC, IIIT, H	Coming up	-
Natural Language Understanding System	N	DESIKA	Sanskrit (plain and accented written text)	C-DAC, B	http://www.tdil.mit.gov.in/download/menu.htm#desika (download)	Windows
Parsing: generation and analysis (parsing)	Y	-	Sanskrit	C-DAC, B	Coming up	-
Parser	N	-	Indian languages	LTRC, IIIT, H	Coming up	-

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Semantics and syntax analyser	Y	Shabdabodha	Sanskrit	ASR, M	http://tdil.mit.gov.in/download/menu1.html	MS-DOS 6.0 or higher with GIST shell
Speaker Verification System	N	-	Indian languages	LTRC, IIIT, H	Coming up	-
Speech Recognition System	N	-	Indian languages	LTRC, IIIT, H	Coming up	-
Speech synthesizer	N	-	Hindi	CEERI, D	Coming up	Sound blaster card with speakers
Spell-checker	Y	ILEAP: Internet ready Indian language word processor	Multilingual (15 IL)	C-DAC, B	http://www.cdacindia.com/html/gist/products/ileap.asp (download)	Windows
Spell-checker	Y	Webdunia Spell Checker	Indian languages	Webdunia	http://www.webdunia.net/products/SpellChecker.asp (product info. only)	Windows
Spell-checker	Y	Anuvaadak	English and Hindi	Super Infosoft	-	Windows family
Spell-checker	N	Akshara-XP	English, Hindi	Aryan Softwares	-	Windows
Spell-checker	N	Su-windows	Hindi	R.K. Compusoft Pvt. Ltd.	-	-
Spell-checker	N	SULIPI 2.0	Hindi	SEACOM	-	-
Spell-checker	N	-	Punjabi	C-DAC, MH	http://www.tdil.mit.gov.in/download/menu.htm	MS-DOS File Size 406 KB approx.

Resource/ Service/ Tool	PoA	Name	Languages	Agency	Availability	System Requirements
Spell-checker	N	Nerpadam	Malayalam	C-DAC, T	http://www.malayalamresourcecentre.org/Mrc/products/nerpadam.html (product info. only)	-
Syntactic and Semantic Analyser	Y	-	Sanskrit	C-DAC, B	Coming up	-
Syntactic and Semantic Analyser	Y	-	Sanskrit	ASR, M	-	DOS platform with GIST card; and is being ported to Windows
Syntactic and Semantic Analyser	Y	-	Sanskrit	C-DAC, B	Coming up	-
Tagger, lemmatiser	Y	-	Sanskrit	C-DAC, B	Coming up	-
(word level) Part of Speech Tagger	N	-	-	TDIL	Coming up	-
Text to Speech Synthesiser	N	-	Telugu and Hindi	LTRC, IIIT, H	http://nlp.iiit.net/~speech/ (demonstration)	-
Thesaurus	Y	Sanskrit Authoring System	Sanskrit	C-DAC, B	Coming up	Windows
Word Processor	N	ILeap: Internet ready Indian language word processor	15 IL	C-DAC, B	http://www.tdil.mit.gov.in/download/menu.htm#ileap (download)	Windows

Brief description of Resources, Services and Tools

Please see the disclaimer near the end of the document.

1. Anglabharati by Indian Institute of Technology, Kanpur

http://www.iitk.ac.in/
http://www.cse.iitk.ac.in/users/langtech/hist.htm
http://www.cse.iitk.ac.in/users/langtech/anglabharti.htm

The system is a machine aided translation system for translation between English to Hindi, for the specific domain of Public Health Campaigns. Anglabharti uses a pseudo-interlingua approach. It analyses English only once and creates an intermediate structure that is almost disambiguated. The intermediate structure is then converted to each Indian language through a process of text-generation. The effort in analyzing the English sentences is about 70% and the text-generation accounts for the rest of the 30%. Thus only with an additional 30% effort, a new English to Indian language translator can be built.

Anglabharti is a pattern directed rule based system with context free grammar like structure for English (source language) that generates a `pseudo-target' applicable to a group of Indian languages (target languages). A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the ‘pseudo-target' are constructed. The idea of using `pseudo-target' is primarily to exploit structural similarity to obtain advantages similar to that of using interlingua approach. It also uses an example-base to identify noun and verb phrases and resolve their ambiguities.

The ANGLABHARTI methodology was used to design a functional prototype for English to Hindi on Sun system. Feasibility on extending this for English to Telugu/Tamil was also demonstrated. AnglaHindi software technology has been transferred to two organizations and is being made available on both the Linux and Windows platforms.

1.2 Spell-checker by Indian Institute of Technology, Kanpur

http://www.cse.iitk.ac.in/users/rmk/proj/proj.html#spell

The approach to design of a spell checker is to develop a user error model for each class of user where the source of error may the due to incorrect phonetics, inaccurate inputting, or other influences. The spell-checker uses this error-model in making suggestions for the error.

2. Anusaaraka by Indian Institute of Technology, Kanpur and University of Hyderabad

http://www.iiit.net/ltrc/
http://www.iiit.net/ltrc/Anusaaraka/anu_home.html
http://www.iiit.net/ltrc/Publications/anu_brief.html

The task of building an MT System is subdivided into two parts: 1. The first module (called core anusaaraka) does language-based analysis: It takes all the information in the source text and presents it in its output, in an intermediate language that is quite close to the target language. 2. The second module may do domain specific knowledge based processing, statistical processing, etc. in which it may utilize world knowledge, frequency information, concordances, etc. to produce output in the target language.

For more, see [1]

3. Matra by Centre for Development of Advanced Computing, Mumbai (erstwhile NCST)

http://www.ncst.ernet.in/matra/
http://www.ncst.ernet.in/matra/about.shtml

MaTra is an ongoing project at C-DAC, Mumbai. It aims at Machine Aided Translation from English to Hindi. Work is going on in news domain, but the approach is applicable for any domain. The system breaks an English sentence into chunks, analyzes the structure and displays the Hindi output. A prototype can translate simple (single-verb-group), assertive sentences. Work is on to increase the range of sentences.

4. MANTRA by Centre for Development of Advanced Computing, Bangalore

http://www.cdacindia.com/html/about/success/mantra.asp

MANTRA (MAchiNe assisted TRAnslation tool) translates English text into Hindi in a specified domain of personal administration, specifically gazette notifications, office orders, office memorandums and circulars. MANTRA uses *Lexicalized Tree Adjoining Grammar (LTAG)* to represent the English as well as the Hindi grammar. It uses *Tree Adjoining Grammar (TAG)* for parsing and generation. It has a modified, Earley's style *bottom-up parsing* algorithm to speed up the parser. It uses several pre-processing tools: phrase marker, domain specific identifiers like proper nouns, dates and other entities, spell-checker, grammar checker. It allows online word addition and grammar creation and updating.

The MANTRA Technology is being expanded to translate English texts into other Indian languages such as Gujarati, Bengali, and Telugu. The domain for Hindi translation is being expanded from the domain of personnel administration to other domains like banking, transportation and agriculture.

C-DAC, B is exploring possibilities in speech recognition and speech synthesis.

4.1 Saranshak by Centre for Development of Advanced Computing, Bangalore

http://www.cdacindia.com/html/aai/saran.asp

Saranshak (The Summarizer) is a natural language based summarizer. It abstracts key content from one or more information sources. Summarizer has two approaches: extraction and abstraction. Extraction involves selecting original pieces from the source document and concatenating them to yield a shorter text. This approach does little to ensure that the summary is coherent, which can make the text hard to read. Abstraction paraphrases in more general terms what the text is about. Currently, this system uses a concept and name based information extraction approach. It uses a set of ranking strategies on sentence and on word level to calculate the relevancy of a sentence to a document. It extracts the most relevant sentences. It creates a summary of the document from these sentences. The user can set the length of the summary.

5. UCSG MAT by University of Hyderabad

http://www.uohyd.ernet.in/

MAT is a machine aided translation system for translating English texts into Kannada. It requires post-editing. It works at sentence level. It parses an input sentence using the UCSG (Universal Clause Structure Grammar) parsing technology (developed by Dr. K. Narayana Murthy) and then translates it into Kannada using the English-Kannada bilingual dictionary, Kannada Morphological Generator and the translation rules.

6. UNL MT by Indian Institute of Technology, Mumbai

http://www.cfilt.iitb.ac.in/

IIT, Mumbai is the Indian participant in Universal Networking Language (UNL) project. UNL is an international project of United Nations University. UNL is an interlingua for semantic representation. Input in the source language is 'enconvrted' into UNL and then 'deconverted' from UNL to the target language. Currently, work on enconversion and deconversion in English, Hindi and Marathi is going on.

For more, refer to [3].

6.1 Marathi and Hindi WordNets by Indian Institute of Technology, Mumbai

http://www.cfilt.iitb.ac.in/wordnet/webmwn/ Marathi WordNet
http://www.cfilt.iitb.ac.in/wordnet/webhwn/ Hindi WordNet

These WordNets are compatible with English WordNet and Euro WordNet. There are 5521 synsets in Marathi WordNet and 11,312 in Hindi WordNet. The work on Konkani will start in collaboration with Goa research group.

7. Tamil Anusaaraka by Anna University's K. B. Chandrasekhar Research Centre, Chennai

http://www.au-kbc.org/frameresearch.html

The aim is to build a Human Aided Machine Translation System for English-Tamil. The MT system has three major components, viz. morphological analyser of source language, mapping unit and the target language generator. The Tamil-Hindi Machine Aided Translation (MAT) system has a performance in the range of 75%. The state-of-the-art Tamil Morphological analyser can handle nearly 3.5 million word forms including compound words with more than 95% accuracy.

7.1 Word Sense Disambiguation (WSD) in Tamil by Anna University's K. B. Chandrasekhar Research Centre, Chennai

http://www.au-kbc.org/research_areas/nlp/projects/wsd1.html

The aim is to reduce the human effort needed for sense tagging. This approach is similar to and an extension of *Context-group discrimination*. All the occurrences of the ambiguous words are classified into different clusters in such a way that all the occurrences are in the same sense within a cluster. Then co-occurrence words are collected for each cluster. These words are used for manually assigning the sense for each cluster. It is planned to probe the applicability of the inflections of words in WSD for rich inflectional languages like Tamil. The hypothesis is that, "Each sense of an ambiguous word will predominantly co-occur with words in a particular inflected form". The preliminary investigations reveal that the hypothesis is indeed useful for some senses of an ambiguous word if not for all senses. Therefore, it is proposed to use this information simultaneously with the co-occurrence information explained earlier. The system uses context-based approach, case relation based approach and integrates the two approaches.

7.2 Biological Named Entity Recognizer by Anna University's K. B. Chandrasekhar Research Centre

http://www.au-kbc.org/research_areas/nlp/projects/named_entity.html

A new named entity extraction module as a part of information extraction system, which is based on a manually developed set of rules that rely heavily upon some crucial lexical information, linguistic constraints of English, and contextual information. This system achieves state of art results in the biological name detection task, which is what many of the current name extraction systems do. It detects chemical names and obtains a high degree of success in recognizing chemicals. It is hoped that this task can help improve the precision of protein name detection as well. A system to automatically extract the interactions among the biological entities is being developed.

7. 3 Tamil WordNet by Anna University's K. B. Chandrasekhar Research Centre, Chennai

http://www.au-kbc.org/frameresearch.html

A Tamil WordNet is being developed in collaboration with Dr. S. Rajendran of Tamil University, Thanjavur. Tamil WordNet relies on Rajendran's (2001) Modern Tamil Thesaurus that is based on Nida's (1975) Componential Analysis of Meaning. This work is available in the electronic form. Tamil vocabulary is classified into four major domains: entities, abstracts, events and relationals based on the part-of-speech categories.

7.4 Various tools by Anna University, Chennai

http://ns.annauniv.edu/rctamil/html/eproj.htm

Anna University's Resource Centre for Indian Language Technology Solutions for Tamil is currently building morphological analyser, morphological generator, automatic tagger, spell checker, grammar checker, parser, text summarizer, word processor, and text to speech for Tamil.

8. MAT by Jadavpur University

http://www.jadavpur.edu/

Jadavpur University at Kolkata has a rule-based English-Hindi MAT. It uses transfer approach. It works for news sentences.

9. Anuvaadak by Super Infosoft

http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak

The Spell-checker is in both English and Hindi. It has an inbuilt thesaurus and grammar checker. Inbuilt grammar checker works in pre-translation and post-translation stages. It has inbuilt *dictionaries for specific domains* e.g. official, formal, agriculture, linguistics, technical, and administrative. An English word processor is inbuilt. When Hindi meaning of the English word is not available in dictionary, facility of transliteration is provided. The software runs on any operating system in the Windows family.

10. Statistical MT by International Business Machines

http://www.research.ibm.com/irl/projects/translation.html

IBM India Research Lab at New Delhi has started work on statistical machine translation between English and Indian Languages. Their work is based on similar work at IBM for other languages.

10.1 Hindi speech recognition system by International Business Machines

http://researchweb.watson.ibm.com/irl/projects/speech/index.html

IBM has Hindi speech recognition system that uses *Acoustic and Language models*. IBM aims to cover more Indian languages and then to build a multilingual speech recognizer for the Indian languages based on a multilingual phone set. It aims to build Hindi speech recognition technology, Hindi speech synthesizer, Audio-Visual Speech Recognition.

11. Oriya Machine Translation System (OMTrans) by Utkal University, Vanivihar

http://www.ilts-utkal.org/omt.htm

In OMTrans, the source language is English and target language is Oriya. It does sense disambiguation using the N-gram model. It has a parser and Oriya Morphological Analyser (OMA), OGC (Oriya Grammar Checker), OSC (Oriya Spell Checker) and OSA (Oriya Semantic Analysis). These modules contribute to OWP (Oriya Word Processor) which facilitates multilingual editing.

11.1 Sanskrit WordNet by Utkal University, Vanivihar

http://www.ilts-utkal.org/orinet.htm

Utkal University is building WordNet for Sanskrit language using the Navya-NyAya Philosophy and Paninian grammar. Besides the standard semantic relations in WordNet, it has etymology and analogy. These play important roles in Navya-NyAya Philosophy. The project has analysed 300 Sanskrit words (200 nominal words and 100 verbal words).

11.2 Oriya WordNet (OriNet) by Utkal University, Vanivihar

http://www.ilts-utkal.org/orinet.htm

The system has two independent modules. One module is developed to write the source files containing the basic lexical data and these files are taken as the input for OriNet system. Lexicographer takes care the major work of this module. Second module is a set of programs by which it accepts the source files, processes it to display for the user and also provides different interface to use other applications. System has been designed using Object-Oriented paradigm according to Oriya language structure with over 1100 lexical entries.

12. Machine Aided Translation by Centre for Development of Advanced Computing, Noida

http://www.cdacnoida.com/nlp.htm

Machine Aided Translation system translates public health related sentences from English to Hindi. It provides post-editing facility. It fuses Paninian framework with modern artificial intelligence techniques to exploit commonality among Indian languages. According to developers, the system achieves 85% correct parsing and about 60% correct translation.

13. Bangla Spell Checker by Indian Statistical Institute, Kolkata

http://www.isical.ac.in/~rc_bangla/products.html

Bangla Spell Checker: works with the help of Bangla Editor. It shows suggestions list for each wrong word. It allows the user to add words to and delete words from the dictionary.

14. Dictionary, Spell-checker and Word-processor by Webdunia

http://www.webdunia.net/products/Dictionary.asp
http://www.webdunia.net/products/SpellChecker.asp

Webdunia has a dictionary that suggests synonyms; a spell-checker for all Indian languages and a word-processor called Windic.

URLs

- **Institute, Organisation**
- **Online services: translation, spell-checking and tagging**
- **Dictionary**
- **Pictorial Glossary, Pictorial Dictionary and Common Vocabulary**
- **Computing Terms, Computing Literature**
- **Others**

Institute, Organisation

Anglabharati (IIT-K, C-DAC, NOIDA)
http://www.iitk.ac.in/
http://www.cse.iitk.ac.in/users/langtech/hist.htm
http://www.cse.iitk.ac.in/users/langtech/anglabharti.htm

Anusaaraka (IIT-K, UoH)
http://www.iiit.net/ltrc/
http://www.iiit.net/ltrc/Anusaaraka/anu_home.html
http://www.iiit.net/ltrc/Publications/anu_brief.html

MaTra (C-DAC, M)
http://www.ncst.ernet.in/
http://www.ncst.ernet.in/./kbcs/nlp.shtml
http://www.ncst.ernet.in/matra/
http://www.ncst.ernet.in/matra/about.shtml

Mantra (C-DAC, B)
http://www.cdac.org.in/
http://www.cdacindia.com/html/aai/mantra.asp

UCSG MAT (UoH)
http://www.uohyd.ernet.in/
http://www.languagetechnologies.ac.in/

UNL MT (IIT-B)
http://www.cfilt.iitb.ac.in/
http://laiir.cse.iitb.ac.in/eng_unl_anal.html
http://www.cfilt.iitb.ac.in/eng-hin-mt/

Tamil Anusaaraka (AU-KBC, C)
http://www.au-kbc.org/

MAT (JadavpurU)
http://www.jadavpur.edu/

Anuvaadak (Super Infosoft)
http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak

StatMT (IBM)
http://www.research.ibm.com/irl/projects/translation.html

a. TDIL: Technology Development in Indian Languages, Ministry of Information Technology, Govt. of India
   http://tdil.mit.gov.in/

b. International Institute of Information Technology, Hyderabad
   http://www.iiit.net/ltrc/

c. Central Institute for Indian Languages, Mysore
   http://www.ciil.org/

d. Utkal University
   http://www.ilts-utkal.org/nlp.htm
   http://www.ilts-utkal.org/speech.htm

e. Resource Centre for Indian Language Technology Solutions Bangla
   http://www.isical.ac.in/~rc_bangla/products.html#corpus

Online services: translation, spell-checking and tagging

a. Translation from English to Hindi by Anglabharati (IIT-K, C-DAC, NOIDA)
   http://anglahindi.iitk.ac.in/index2.html
   http://anglahindi.iitk.ac.in/newpages/footer.htm

b. Anuvaadak, a English-Hindi translation software
   http://www.mysmartschool.com/pls/portal/portal.MSSStatic.ProductAnuvaadak

c. Shakti (Version 0.58) - IIIT-Hyderabad Machine Translation System (Experimental)
   http://216.236.98.137/~shakti/

d. Nerpadam - Malayalam spell-checker
   http://www.malayalamresourcecentre.org/Mrc/products/nerpadam.html

e. Webdunia spell-checker
   http://www.webdunia.net/products/SpellChecker.asp

f. Hindi morphological tagger
   http://ccat.sas.upenn.edu/plc/tamilweb/hindi.html

Dictionary

a. Trilingual dictionary (English, Hindi and Malayalam)
http://www.malayalamresourcecentre.org/Mrc/products/trilingual.html

This is the first online trilingual dictionary for English, Hindi and Malayalam. It contains more than 50,000 words in each language. It includes idioms, glossary of foreign words, usages in English, Hindi and Malayalam, etc. Search on a word in any of the three languages gives the meaning in the other two languages. It allows searching on parts of speech, prefix, and suffix. Prepared by Resource Centre for Indian Language Technology Solutions – Malayalam.

b. English to Hindi dictionary
http://sanskrit.gde.to/hindi/dict/eng-hin-itrans.html

c. English to Hindi dictionary in three different encodings: ISCII 8 bit, CSX and ITRANS
http://www.archaka.com/puja/english_to_hindi_dictionary1.htm

d. Hindi dictionary and lookup
http://www3.aa.tufs.ac.jp/~kmach/hnd_la-e.htm#wordanalysis
http://www.foreignword.com/Langlinks/Hindi.htm

e. Marathi Dictionary
http://sanskrit.gde.to/all_txt/marathi-dict.txt

f. Dictionaries of Sanskrit
http://www.mavicanet.ru/directory/eng/14377.html

g. Apte Sanskrit Dictionary Search
http://iiasnt.leidenuniv.nl/cgibin/startq.cgi?flags=endnnnl&root=leiden&basename=%5Cdata%5Cie%5Cconcord

h. Capeller's Sanskrit-English Dictionary
http://www.uni-koeln.de/phil-fak/indologie/tamil/cap_search.html

i. Tamil Lexicon - Manual
http://www.uni-koeln.de/phil-fak/indologie/tamil/otl.html

j. Indo-Iranian languages
http://www.yourdictionary.com/languages/indoiran.html

k. List of online dictionaries
http://www.foreignword.com/Tools/transnow.asp?p=files/f_source.htm

Pictorial Glossary, Pictorial Dictionary and Common Vocabulary

a. Pictorial glossary for Bengali
http://www.anukriti.net/dicbooks/pict-bengali/1.html

b. Common vocabulary for Hindi-Kashmiri
http://www.anukriti.net/dicbooks/hindi-kashmiri/1.htm

c. Pictorial glossaries, dictionaries and common vocabularies for some Indian languages
http://www.anukriti.net/tools.asp

d. Tamil Picture dictionary
http://ns.annauniv.edu/rctamil/html/picdic.htm

Computing Term, Computing Literature

a. The natural language group at Information Science Institute
http://www.isi.edu/natural-language
http://www.isi.edu/natural-language/mteval/

b. Online computing dictionary
http://www.instantweb.com/foldoc/source.html

c. Machine Translation: An Introductory guide
http://clwww.essex.ac.uk/MTbook/HTML/book.html
http://www.essex.ac.uk/linguistics/clmt/MTbook/

d. 'Compendium of Translation Software' (555kb), ed. John Hutchins. Sixth edition, March 2003. Paid download:
http://www.eamt.org/compendium.html

e. Free online dictionary of computing
http://foldoc.doc.ic.ac.uk/foldoc/

f. Computer-based translation systems and tools - John Hutchins
http://www.eamt.org/archive/hutchins_intro.html

g. Machine Translation: past, present, future
http://ourworld.compuserve.com/homepages/WJHutchins/PPF-TOC.htm

Others

a. Sanskrit resources
   http://sanskrit.gde.to/

b. International languages translation resources
   http://www.foreignword.com/technology/other/other.htm

c. In-depth information about the major languages of the world
   http://www.worldlanguage.com/Languages/

d. World Language page for Marathi
   http://www.worldlanguage.com/Languages/Marathi.htm

e. Rgvedic word concordance
   http://aa2411s.aa.tufs.ac.jp/~tjun/sktdic/

f. Meanings of words
   http://www.wordanywhere.com/

g. Indian language family
   http://www.ciil.org/languages/map4.html

h. Indian Lexicon; semantic and alphabetic sequences of lexemes in Indian languages
   http://www.hindunet.org/hindu_history/sarasvati/html/indlexmain.htm

i. Indian Language Family pie chart
   http://www.ciil.org/languages/map4.html

Glossary of terms

case marker	a marker that indicates the semantic relationship between the predicate and its argument
Interlingua	an intermediate language used for semantic representation common to more than one language
StatMT	Statistical Machine Translation
sub-language	vocabulary and grammar of a particular subject field
synset	synonym set – a set of words that point to a unique concept
Transfer/Frames	a transfer method that uses frames
Transfer/Rules	a transfer method that uses rules
Transfer/UCSG	a transfer method that uses UCSG
Transfer/XTAG	a transfer method that uses XTAG
UCSG	Universal Clause Structure Grammar
UNL	Universal Networking Language
WordNet	a lexical knowledge base of semantic relations (synonymy, antonymy, hypernymy, hyponymy, holonymy and meronymy). It is compatible with English WordNet and Euro WordNet. http://www.cogsci.princeton.edu/~wn/
WSD	Word Sense Disambiguation

Bibliography

1. Bharati, Akshar, Chaitanya, Vineet, Kulkarni, Amba P., Sangal, Rajeev "Anusaaraka: Machine Translation in stages". Vivek, A Quarterly in Artificial Intelligence, Vol. 10, No. 3 (July 1997), NCST, India, pp. 22-25.
   http://arxiv.org/pdf/cs.CL/0306130

2. Dash, Niladri Sekhar, Chaudhuri, Bidyut Baran 2000. "Why do we need to develop corpora in Indian languages?". A paper presented at SCALLA 2001 conference, Bangalore.
   http://www.elda.fr/proj/scalla/SCALLA2001/SCALLA2001Dash.pdf

3. Dave, Shachi, Parikh, Jignashu and Bhattacharyya, Pushpak Interlingua Based English Hindi Machine Translation and Language Divergence, Journal of Machine Translation, Volume 17, September, 2002.

4. Hutchins, W. John, Somers, Harold L. An Introduction to Machine Translation. Academic Press, London, 1992.

5. Murthy, B. K., Deshpande, W. R. 1998. "Language technology in India: past, present and future".
   http://www.cicc.or.jp/english/hyoujyunka/mlit3/7-12.html

6. Rao, Durgesh 2001. "Machine Translation in India: A Brief Survey". SCALLA 2001 conference, Bangalore.
   http://www.elda.fr/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf

Disclaimer

This document is based on the information available to the author, and is believed to be accurate as on 31/08/2003, to the best of author's knowledge and belief. No legal claim is made regarding the accuracy of the information.

Acknowledgements

The author thanks the following people:

a. Dr. Pushpak Bhattacharyya for motivation and guidance.
b. Durgesh Rao for a table and contact person's email addresses from his paper titled 'Machine Translation in India: A Brief Survey'.
c. Niladri Sekhar Dash and Bidyut Baran Chaudhuri for a table given in their report 'Why do we need to develop corpora in Indian languages?'.