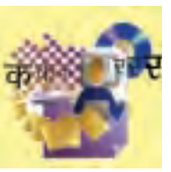

## 5.6 Translation Support Systems (TSS) Contents

### a) Machine Aided Translation / Machine Translation

#### 5.6.1 Anusaaraka

Bharati, Akshar, Amba P Kulkarni, Vineet Chaitanya, Rajeev Sangal, G.Umamaheshwara Rao, *Overcoming the Language Barrier in India, In Translation, Text and Theory, Sage Publishers, New Delhi, 2002.*

**Abstract**

The anusaaraka system makes text in one Indian language accessible in another Indian language. In the anusaaraka approach, the load is so divided between man and computer that the language load is taken by the machine, and the interpretation of the text is left to the man. The machine presents an image of the source text in a language close to the target language. In the image, some constructions of the source language (which do not have equivalents) spill over to the output. Some special notation is also devised. The user after some training learns to read and understand the output. Because the Indian languages are close, the learning time of the output language is short, and is expected to be around 2 weeks.

The output can also be post-edited by a trained user to make it grammatically correct in the target language. Style can also be changed, if necessary. Thus, in this scenario, it can function as a human assisted translation system.

Currently, anusaarakas are being built from Telugu, Kannada, Marathi, Bengali and Punjabi to Hindi. They can be built for all Indian languages in the near future. Everybody must pitch in to build such systems connecting all Indian languages, using the free software model.

#### 5.6.2 Language Access

Bharati, Akshar, Amba P Kulkarni, Vineet Chaitanya, Rajeev Sangal, *An Information Based Approach, Knowledge-Based Computer Systems, Tata McGraw-Hill, New Delhi, Dec. 2000.*

**Abstract**

The anusaaraka system (a kind of machine translation system ) makes text in one Indian language accessible through another Indian language. The machine presents an image of the source text in a language close to the target language. In the image, some constructions of the source language (which do not have equivalents in the target language) spill over to the output. Some special notation is also devised.

Anusaarakas have been built from five pairs of languages: Telugu, Kannada, Marathi, Bengali and Punjabi to Hindi. They are available for use through Email servers.

Anusaarkas follows the principle of substitutability and reversibility of strings produced. This implies preservation of information while going from a source language to a target language.

For narrow subject areas, specialized modules can be built by putting subject domain knowledge into the system, which produce good quality grammatical output. However, it should be remembered, that such modules will work only in narrow areas, and will sometimes go wrong. In such a situation, anusaaraka output will still remain useful.

#### 5.6.3 Machine Translation Activities In India

Bharati, Akshar, Rajeev Sangal, Dipti M Sharma, Amba P Kulkarni *A survey, Appeared in the proceedings of workshop on survey on Research and Development of Machine Translation in Asian Countries, Thailand, May 13-14, 2002*

**Abstract**

This paper outlines the major MT related activities that are being carried out in India. The focus of activities is on building lexical resources for Indian languages and development of machine translation systems. A brief report of activities in speech processing systems, OCRs and knowledge extraction is also presented.

#### 5.6.4 Interlingua Based English Hindi Machine Translation And Language Divergence

Shachi Dave, Jignashu Parikh and Pushpak Bhattacharya, *Journal of Machine Translation, Volume 17, 2002.*

**Abstract**

This work studies the phenomenon of various language divergences that exist between English and Hindi and their implication on the automatic translatability between these languages. The framework of translation is interlingua based and makes use of a recently proposed interlingua called the “Universal Networking Language”. It is shown that the power of the inter- lingua shields the translation process from many of the language divergence caused problems.

#### 5.6.5 Word Sense Disambiguation For Machine Translation

S. Mohanty, R.C. Balabantaray & P. K. Santi, *Published at: CIT-2001, School of Mathematics Statistics and Computer Science, Utkal University, Bhubaneswar, Orissa.*

**Abstract**

Language is meant for expressing feelings of persons. Machine Translation (MT) System is a step towards solving the language problem among different people, which is capable of translating one language to other, so that the system can become more potential for exchange of knowledge.

Several words have various meaning depending on the context in which they are used. Therefore the words are potentially ambiguous, so Natural language

understanding must cope up with various ambiguities. The main task of this problem (i.e. to identify the proper meaning of a word) is to determine which of the senses of an ambiguous word is invoked in a particular use of the word. Looking at the context of the word in which it is used various methods proposed to solve this problem are:

1. Supervised Disambiguation.
2. Dictionary based Disambiguation.
3. Unsupervised Disambiguation.

In Supervised learning it is quiet difficult to train the system for different senses of a word, which is a time consuming process. In Dictionary based it is not always possible to identify the appropriate sense.

Here is a trial to overcome the problems faced in case of Supervised and Dictionary based by using Unsupervised Disambiguation. To assign appropriate meaning (sense) to a given word in a text or discourse, an Unsupervised Learning Algorithm based on use of a word in the respective language are been applied.

This algorithm for solving Word Sense Disambiguation (WSD) can be used in Machine Translation system for proper assignment of meaning to words in a sentence. This helps to overcome one of the problems of realistic Machine Translation.

### 5.6.6 मशीनी – अनुवाद के परिप्रेक्ष्य में तुलनात्मक भाषा – विज्ञान (COMPARATIVE LINGUISTICS) की उपादेयता

सत्यनारायण पाण्डेय,* नारायण गोपाल डोंगरे,** कौशल कुमार शुक्ल * *Paper presented in Symposium on Translation Support System (STRANS) 2002 I.I.T. Kanpur, U.P., held on March 15-17, 2002.*

सारांश

सूचना प्रौद्योगिकी के विकास ने भारत जैसे विकासशील देश के लिये अनेक नई संभावनाओं को जन्म दिया है। साथ ही साथ एक जोखिम (risk) भी सामने आ खड़ी हुई है। औद्योगिक क्रान्ति के समय हमने यह भूल की है कि उसके लाभों को एक छोटे अभिजात्य वर्ग तक ही सीमित कर दिया, वैसी ही स्थिति की पुनरावृत्ति सूचना प्रौद्योगिकी के कारण आंकिक प्रविधि (Digital Divide) के रूप में सामने आ सकती है। जन सामान्य तक सूचना प्रौद्योगिकी के लाभों को पहुँचाने में मुख्य बाधा हार्डवेयर (hardware) की उपलब्धता की नहीं, अपितु भाषायी बाधा (language barrier) है। इसी दृष्टि से मशीनी अनुवाद वह मूल–भूत आवश्यकता है जो इण्टरनेट द्वारा उपलब्ध ज्ञान के महासागर को सामान्य जनता के दैनिक जीवन में उपयोगी बना सकता है। प्रस्तुत लेख में मशीनी–अनुवाद की परिवर्तक (Transfer) प्रणाली [1] पर आधारित साफ्टवेयर (software) बनाने हेतु आवश्यक अंग्रेजी एवं भारतीय भाषाओं के व्याकरण का एक तुलनात्मक विश्लेषण प्रस्तुत किया गया है। यह अध्ययन उन साफ्टवेयर अभियंताओं के लिये उपयोगी साबित होगा जो परिवर्तक प्रणाली (transfer approach) के आधार पर मशीनी–अनुवाद के एक विधा की रचना कर रहे हैं।

आई0 बी0 एम0 (IBM) के अध्ययन के अनुसार विश्व व्याप्त तंत्र (word wide web) पर उपलब्ध जानकारी का लगभग 80% भाग अंग्रेजी में ही उपलब्ध है, जबकि शेष 20% में विश्व की सभी भाषाएँ सम्मिलित हैं। भारतीय भाषाओं की नई तंत्रिका स्थली (web sites) अथवा आश्रय स्थली (portals) बनाकर जानकारी को उपलब्ध कराने की अपेक्षा यह अधिक समीचीन प्रतीत होता है कि मशीनी–अनुवाद द्वारा अंग्रेजी भाषा में उपलब्ध तंत्रिका पृष्ठों (web pages) को स्थानीय भारतीय भाषाओं में दर्शाया जाय। इससे तेजी से बढ़ती हुई नई अंग्रेजी तंत्रिका स्थलों (web sites) की जानकारी भी स्थानीय भाषाओं में उपलब्ध होती रहेगी।

### 5.6.7 Dealing With Unknown Lexicons In Machine Translation From English To Hindi

R.M.K. Sinha, *Proc. Of IASTED International Conference on Artificial Intelligence and Soft Computing, May 21-24, 2001, Cancum, Mexico, pp 333-336.*

**Abstract**

A natural text for translation using machine contains several unknown words for which there are no entries in the dictionary. These words may be names, acronyms, abbreviations, terminologies and foreign words. Also, some of the words may not be found in the dictionary due to its limited size. A machine translation system has to provide mechanism for handling such unknowns. In this paper we describe the strategy adopted in our system for machine aided translation from English to Hindi. No attempt has been made to expand the vocabulary by deriving their meaning. Instead, a transliteration in Hindi with appropriate suffixes or appendage is used to substitute for their meaning. It is a common practice in India to mix the words of English in Hindi and vice-versa. However, the grammatical rules in construction of gender, number, verb-nominalization or forms, conform to that for the language(Hindi or English) used irrespective of their origin. It is found that the number of alternate translations generated is directly proportional to the number of unknowns which are not found in the dictionary. However, the acceptable translation is invariably contained in the multiple translations generated. We use a number of heuristics to identify the type of unknown and limit the number of alternatives.

### 5.6.8 Translating News Headings From English To Hindi

R.M.K. Sinha, *6 IASTED International Conference on Artificial Intelligence and Soft Computing (AC2002), Banff, Canada, July 17-19, 2002.*

**Abstract**

A news heading is meant to convey the primary focus of the news item. As it is in an abstracted form, it may have a grammar of its own. Many a time, it uses an oblique/inflected form of tense and narration. This obliqueness may be different in different languages. For example, an English heading “Hundreds die in flood” is actually used to convey “Hundreds died in flood” denoting occurrence of an event. It is this past tense form that should get translated into the target language and not the present tense form of the English source. However, a heading such as “Hundreds die in flood every year”, should remain the same for

translation as it reflects continuity of an event. In this paper, I have considered the problem of translating news headings from English to Hindi. The strategy consists of categorizing English news headings into a number of classes and providing transformation rules for obtaining the expanded form of the heading. The expanded form of the heading is then translated into Hindi and then skimmed to yield the corresponding news heading in Hindi.

### 5.6.9 VAASAANUBAADA Automatic Machine Translation Of Bilingual Bengali-Assamese News Texts

Kommaluri Vijayanand, S I Choudhury, Pranab Ratna, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

#### Abstract

This paper presents a project to translate the bilingual Bengali-Assamese news texts using Example-Based Machine Translation technique. The work involves machine translation of bilingual texts at sentence level. In addition, the work also includes preprocessing and post-processing tasks. The work has its own uniqueness because of the language pair that is chosen for experimentation. We constructed and aligned the bilingual corpus manually by feeding the real examples using pseudo code. The longer input sentence is fragmented at punctuations, which resulted in high quality translation. Backtracking is used when the exact match is not found at the sentence/fragment level, leading to further fragmentation of the sentence. Since, bilingual Bengali-Assamese languages belongs to the Magadha Prakrit group, the grammatical form of the sentences are very similar and have no lexical word groups. The results when tested are fascinating with quality translation.

### 5.6.10 Computational Relational Lexicon For Machine Translation

Om Vikas, *International Conference on Natural Language Processing, Kyushu Institute of Technology 1993*

#### Abstract

Lexicon plays an important role in natural language processing. For each lexical entry, it contains definition, lexical and semantic relations. Lexical relations include synonyms, antonyms, co-locations, special situation words, magnification, group, symptoms, etc. Entities and Actions are two types of concepts Semantic relations include Entity-Entity, Entity-Action, Action-Action relations and features of Entity and Actions. 14 basic acts are proposed similar to these proposed by Roger Schank. But these are somewhat different. Relational lexicon has been described in the context of machine translation.

### 5.6.11 मशीनी अनुवाद की समस्याएं (Problems in Machine Translation)

ओम विकास *अनुवाद विज्ञान : सिद्धांत और अनुप्रयोग*, (सं. डॉ नगेन्द्र), *हिन्दी माध्यम कार्यान्वय निदेशालय, दिल्ली वि. वि. 1993, पृष्ठ 338 से 376.*

#### सारांश

ज्ञान बढता ही जा रहा है। यह बढ़ोतरी विविध भाषा-भाषी समुदायो के द्वारा किए जा रहे नव-नवीन प्रयोगों का परिणाम है। एक देश में भी कई स्वीकृत व्यावहारिक भाषाएँ मिलती हैं। इन कारणों से अनुवाद की आवश्यकता बढ़ती जा रही है। विज्ञान और टैक्नोलॉजी के क्षेत्र में भाषाई जटिलताएँ - अनेकार्थता और लक्षणार्थकता-कम हैं, जबकि सामग्री बहुत अधिक है। ऐसे क्षेत्रो में मशीनी अनुवाद सफल होने की संभावनाएँ अधिक हैं। मानव अनुवाद और मशीनी-अनुवाद के मध्य मानव-मशीनी सहकार उपयोगी सिद्ध होगा। शब्दगत विश्लेषण के साथ-साथ अर्थगत संबंधों को जोड़कर ही संकल्पनाओं के अंतःनिरूपण को स्वीकार्य पाठ में संश्लेषित किया जा सकता है। इसलिए संबंधपरक शब्द-संहिता प्रस्तावित है। इसमें शब्द की परिभाषा, प्रयोग और व्याकरणिक अभिलक्षणों के साथ शब्दगत तथा अर्थगत संबंधों का भी उल्लेख होता है। कुछ व्यावहारिक मशीनी अनुवाद प्रणालियों के प्रणाली-डिजाइन और इनमें मानवीय संपादन की सीमा की तुलना भी की गई है। अंत में, मशीनी अनुवाद के क्षेत्र में शोध की नई दिशाओं-व्याकरणिक थ्योरी, अंतरभाषा, शब्द संहिता, मूल्यांकन और कृत्रिम-बुद्धि तकनीकों का भी उल्लेख किया गया है।

## b) Universal Networking Language

### 5.6.12 Language Independent Natural Language Generation From Universal Networking Language

Hrishikesh Bokil and P. Bhattacharyya, *Second Symposium on Translation Support Systems (STRANS), IIT Kanpur, India, March, 2002.*

#### Abstract

In this work we have built a language independent generator which was tested on a real life corpus for Marathi generation. The whole engine has been built using the object oriented paradigm. Complete syntax planning for Marathi has been investigated rigorously.

### 5.6.13 Multilingual Information Processing Using Universal Networking Language

P. Bhattacharyya, *in Indo UK Workshop on Language Engineering for South Asian Languages (LESAL, Mumbai, India, April, 2001).*

#### Abstract

This paper discusses a multiway information transfer system among the languages English, Hindi and Marathi. The natural language sentences are analysed into the Universal Networking Language and then language specific generators produce sentences from these expressions. The relative complexity of processing these languages is also investigated.

### 5.6.14 Generation of Hindi Sentences From Semantic Structures

Vijay Dwivedi, *M.Tech thesis, Department of Computer Science and Engineering, IIT Bombay, 2002.*


### Abstract

This project reports the experience of generating Hindi sentences from actual corpora of UNL expressions. The corpus is called "The Barcelona Corpus" in which UNL expressions have been manually generated from documents of 4 languages- namely,Spanish, French, Italian and Russian. These expressions were then converted to Hindi. The study shows the effectiveness of using the UNL as a vehicle of language independent representation.


#### 5.6.15 Knowledge Extraction From Hindi Texts

Shachi Dave and P. Bhattacharyya, *Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4, July, 2001.*


### Abstract

This work gives a method of analyzing Hindi sentences into a set of Universal Networking Language expressions. The expressions essentially define a semantic network like structure and hence can be looked upon as the knowledge extracted from the sentences. The analysis relies on "predicate preserving" and addresses complex language phenomena of Hindi.


#### 5.6.16 Natural Language Generation From Semantic Net Like Structures With Application To Hindi

Rayner D'Souza, G.Shivakumar, D.Swathi, P. Bhattacharyya, *in First International Symposium on Translation Support Systems, IIT Kanpur, India, February, 2001.*


### Abstract

We present the design of a natural language generation system that takes the analysed structure of compound-complex sentences as input and generates Hindi sentence as output. Generation of compound-complex sentences for English-Hindi Machine translation is a challenging task, because of the linguistic gap between the two languages. The task requires designing of Structural representation, which can encode the information contained in the English sentence, and designing a transfer module, which can generate the Hindi sentence from the structural representation, with the use of rule sets and the lexicon. The rule sets are used to fill in the gap between the two languages. Rule sets include rules for mapping prepositions, verbs, inflection of nouns, etc. The lexicon is used to get the root form of the Hindi equivalent of the English word.


## c) Wordnet

#### 5.6.17 Indian Logic Based Conceptual Ontology Using Description Logics

G.Aghila, Dr.Ranjani Parthasarathy and Dr. T.V. Geetha, *National Conference on Document Analysis and Recognition, Mandya, Karnataka – 13th and 14th July 2001.*


### Abstract

Language understanding still remains a challenging task. An effective representation of world knowledge is required for natural language processing/computing. Among the many approaches for the design of representation for world knowledge available, we have chosen ontology. The conceptual ontology aids in representing concepts/ entities and other world knowledge in a language independent way. In this paper, the classification of world knowledge used for the design of conceptual ontology is the classification scheme of Indian Logic system-TARKA/NYAYA shastra. This framework attempts to classify all entities in the world from atom to universe into seven basic categories and each category is further divided into subcategories. In addition to this, Nyaya shastra defines relations between concepts/entities in a hierarchical fashion. In this work some of the relations are considered. We have designed special general/primitive constructors for Description Logic(DL), where DL is a powerful class of logic based knowledge representation language. The relations we have considered here are Part-Whole, Generality, Inherence, Contact-action, Contact-contact, Absence–temporal, Presence–temporal, Cause-effect, Limit, Pervade, Qualify, Use, Determinant and Absence- environment. The relation can be between concepts/entities at any level in the knowledge hierarchy. This wide range of relations possible between concepts/entities give us a better understanding of world knowledge. This conceptual ontology can be used by a natural language understanding system to understand concepts conveyed by sentences from a pragmatic viewpoint.


#### 5.6.18 Tamil Wordnet

Devi Poongulhali. P., Kavitha Noel. N., Preeda Lakshmi.R., A. Manavazhahan And T.V. Geetha, *1st International Global Wordnet Conference, CIIL, Mysore, Jan 21-25,2002.*


### Abstract

This paper on 'Tamil Wordnet' presents the design and implementation issues involved in creating a lexical database for Tamil language. The infrastructure of the Tamil Wordnet differs from its standard prototypes, to accommodate the unique features and specialties that are characteristic of Tamil language. The linguistic aspects of Tamil dictate the design of the Wordnet. The implementation, details such as the design of the lexicographic files, database tables, grinder utility, etc have been discussed in the course of the paper. An application to demonstrate the use of Tamil Wordnet has also been looked up on.

### 5.6.19 An Object Oriented Design Approach To Orinet System: On Line Lexical Database For Oriya Language

Sanghamitra Mohanty & Prabhat Kumar Santi, *International Conference on LEC,2002, Hyderabad.*

#### Abstract

One of the major problems in the implementation of Natural Language Processing (NLP) or Machine translation (MT) is a complete lexicon: the place where the systems information about words is stored. There are difficulties in deciding what information should be stored in a lexicon and even greater difficulties in acquiring this information in proper form. OriNet system designed to incorporate multiple lexical database and tools under one consistent functional interface in order to facilitate systems requiring syntactic, semantic and lexical information of Oriya language. We divide the whole work into two independent task. One task is to write the source file that contains the basic lexical data and the content of those files are the lexical substance of OriNet. Lexicographer did the major work of this task. In the second task was to create a set of programs those would accept the source files and processing it ultimately to display for the user. This paper describes an ongoing work on designing an Object Oriented model for OriNet system. The technology of Object Oriented programming in particular the rich library of classes and programming principles in which Java offers. It also provides a convenient tool to conceptualize the process of OriNet system. This technique also allows flexibility and extensibility of the system with more robustness.

### 5.6.20 Making Of A Sanskrit Word-Net

S. Mohanty, K.P. Das Adhikary, P.K. Santi and S. N. Nayak*, *International Conference on Universal Knowledge and Languages, 2002, Goa*.

#### Abstract

Sanskrit language is the base for most of the Indian languages. It has link with foreign languages too. To study this language and also to use it for knowledge enhancement, effective Machine Translation (MT) from one language to the other is necessary for which on line lexical resources are needed. Towards making such resources we have tried to develop a Word-Net for Sanskrit language using the *NavyanyAya* philosophy and *Paninian* grammar. The architecture of this "Sanskrit Word-Net" is described in the paper. Besides Synonymy, Antonymy, Hypernymy, Hyponymy, Holonymy and Meronymy we have also discussed on Etymology and Analogy separately as they play important roles in *NavyanyAya* philosophy which is the specialty on this Sanskrit Word-Net for a better classification of words.

### 5.6.21 Oriya Word Net

S.Mohanty & R.C.Balabantaray & P.K.Sant, *Published at: Proceedings of the 1st Global WordNet Conference , 21-25 January 2002, CIIL, Maysore.*

#### Abstract

Machine Translation (MT) in Oriya language is in its infancy. Nonavailability of proper Electronic dictionary has handicapped us to tackle the MT problem. This inspired us to develop the WordNet in Oriya language. We have tried to design the WordNet taking into account the speciality of this language too. In this WordNet the behavior of each word and its category are being explained.

### 5.6.22 An Experience In Building The Indo Wordnet- A Wordnet For Hindi

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande and P. Bhattacharyya, *First International Conference on Global WordNet, Mysore, India, January, 2002.*

#### Abstract

Wordnets are now being recognized as the essential lexical resource for any NLP and MT systems. In this work, we present our experiences on building the wordnet for Hindi. The synsets are constructed as fundamental entities by consulting various dictionaries and thesauri and then they are linked by the semantic relations of Hypernymy, meronymy,antonymy etc. The size of the wordnet now is about 10,000 synsets. The Hindi wordnet will be linked with the other language wordnets of India and also with the Euro Wordnet and the Princeton wordnet. The system has sophisticated user and data entry interfaces. The underlying database is the mysql database platform.

### 5.6.23 Building Large Scale Ontology Networks

Vasudeva Verma, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

#### Abstract

Adoptable, high performing, large scale ontologies that can be extended to support multi-media play a crucial role in building effective content and knowledge management systems and applications. In the context of developing a Unified Taxonomy and Ontology Network (UTON), we have undertaken the task of developing a technology framework for building large scale ontologies. This paper describes the architecture of UTON and how general purpose resources such as WordNet and open directory project can be used in creating large scale ontology networks and how application specific taxonomies or ontologies can be derived from these general purpose tools and resources.