Development of Resources and Techniques for Processing of some Indian Languages

(A Glimpse of SNLP Activities in India)

Shyam S Agrawal

Advisor,CDAC,Noida and Executive Director,KIIT,Gurgaon

Email: ss_agrawal@hotmail.com, ssagrawal@cdacnoida.in

Invited lecture –LDC,Upenn,PA July 17,2008

Objective-To present an overview of :

- CDAC,Noida –Major Areas of research
- Indian Languages-Some Important properties
- Brief review of the work done for Indian Languages in Development of Resources(Text and speech Corpora etc.)
- Development of specialized Tools/Techniques for processingText/Speech corpora
- Details of ELDA –Hindi Corpus
- Details of CFSL-Speaker Identification Database for Forensic Applications
- Brief about A-Star Project

C-DAC, NOIDA UNIT

NATURAL LANGUAGE PROCESSING AND INTERFACES

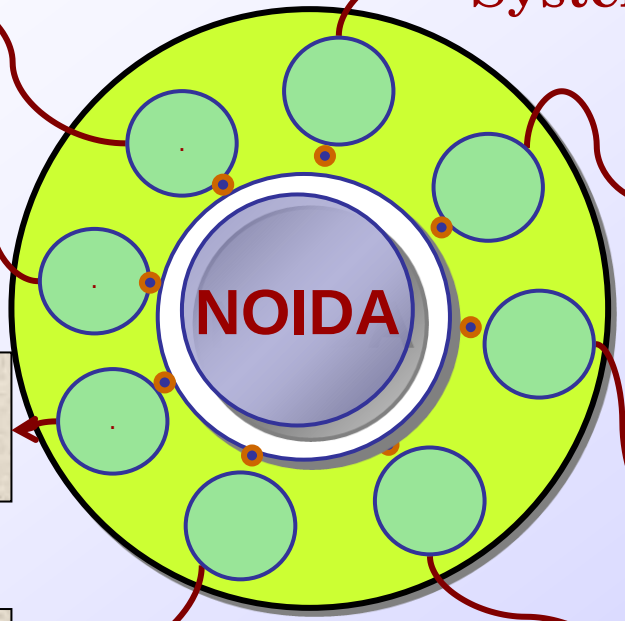
INFRASTRUCTURE AND SUPPORT SERVICES

MISSION C-DAC

HUMAN RESOURCE DEVELOPMENT IN HITECH AREAS

SPECIAL INDUSTRIAL APPLICATIONS

AREAS OF COMPETENCE



- NLP
- Graphical Display System
- E-Governance
- Security Systems
- Internet on CATV & E-Commerce
- Embedded System
- Solar Energy System
- System Engineering and Consultancy

Areas of Expertise in NLP

- Translation System
- Optical Character Recognition
- Text Processing
- Speech Technologies
- Tools development.
- Content Creation
- Web Technology

Indian Languages – Some properties

- ***Many Languages***
- A variety of scripts, and hundreds of dialects
- Eighth Schedule, lists twenty two Scheduled Languages - ***Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu, Bodo, Dogri, Maithili and Santhali***
- Hindi is spoken by 43% population of India followed by Bengali,Telugu,Marathi and others

Language Map

- *Punjabi/Urdu is also spoken by the natives of Pakistan which is much influenced by Persio-Arabic.*

Indian Language properties

- Scripts used are phonetic in Nature
- Better Articulatory discipline
- Systematic manner of production
- Five or Six distinct places of Articulation
- Various types of Flaps/Taps or Trills
- Fewer fricatives compared to English / European languages
- Presence of retroflex consonants
- A significant amount of vocabulary in Sanskrit with Dravidian or Austroasiatic origin gives indications of mutual borrowing and counter influences

Articulatory Classification of Hindi and English Consonants

POA / MOA		Bilabials		Dentals: Labio Dentals		Dentals: Lingua Dental		Alveolar		Retroflex		Palatal/ Paleto Alveolar		Velar		Glottal	
Stop & Affricates	UvUa	p	प				त	t			ट	ॾ	च	k क़	क		
	VoUa	b	ब				द	d			ड	ॻ	ज	ɡ	ग		
	UvAs		फ				थ				ठ		छ		ख		
	VoAs		भ				ध				ढ		झ		घ		
Fricative	Uv	फ़		f		θ		s	स		ष	ʃ	श	ख़		h	ह
	Vo			v		ð		z				ज़ ž		ग़			
Vowel like(Approximant)		w	व					r l	र ल		ड़ ढ़	Y	य				
Nasal		m	म					n	न		ण		ञ	ŋ	ङ		

	Hindi	English
Plosives	16	6
Affricates	4	2
Fricative	4	9
Vowel like (Approximants)	6	4
Nasals	5	3
Total	35	24

MOA=Manner of Articulation
VoUa=Voiced Unaspiration
POA=Place of Articulation

UvUa=Unvoiced Unaspirated
VoAs=Voiced Aspirated
UvAs=Unvoiced Aspirated

Dr. S. S. AGRAWAL

Indian Language properties: Some Exceptions

- In Tamil language, all plosives of a given place of articulation are represented by a single grapheme. The pronunciation of such graphemes depend on the context.
- More fricative consonants are present in Hindi, Punjabi, Sindhi, Kashmiri and Urdu due to influence of Perso-Arabic & English
- च and ज are dental-alveolar in Marathi only, while these are alveolar in Hindi
- ड़ and ढ़ are present in Hindi, Urdu, Sindhi, Punjabi & Oriya.
- Fricative स is श or ष in Oriya
- व and ब are pronounced as ब in Bengali

Indian Language properties: Some Exceptions

- व and ब are pronounced as भ in Oriya
- Punjabi is a tonal language mostly in aspirated voiced sounds
- Sindhi language has implosives
- Native words of the Dravidian languages do not contain aspirated sounds.
- ण sound is more frequently used in Gujarati and Marathi
- त and द are pronounced mostly as ट and ड in Assamese and Gujarati languages
- य and स are pronounced as ज and ह in Assamese

Current scenario

- Text & Speech Corpora
- Machine Translation
- Text to Speech Synthesis
- Speech Recognition
- Tools for Text & Speech processing

Speech /NLP Activities in different institutions

Institute / Languages	Activities
CEERI Delhi (Hindi, Bangali)	•Speech Recognition •Speech Synthesis •Speech /Data bases
TIFR Mumbai (Hindi, Bengali, Marathi, Indian English)	•Speech Recognition •Speech Synthesis •Language Modeling •Speech Data Bases
IIT Kanpur (Hindi, Urdu)	•Machine Translation •Speech Recognition
IIIT Hyderabad (Hindi, Telugu other languages)	•Machine Translation •Speech synthesis •Corpora/Databases
University of Hyderabad (Hindi, Telugu, Urdu other languages)	•Language Identification •Text Corpora Annotation •Machine Translation
IIT, Chennai (Tamil, Hindi)	•Speech Recognition •Speech Synthesis
IIT Mumbai (Hindi, Marathi)	•Machine Translation •Speech Processing
IIT Kharagpur (Hindi, Bengali)	•Machine Translation •Speech Synthesis
IISc Bangalore (Hindi, Tamil etc.)	•Machine Translation •Speech Recognition
CDAC Pune (Hindi, Indian English, other languages.)	•Machine Translation •Speech Recognition •Speech Synthesis
CDAC Noida (Hindi, Punjabi, Marathi, other languages)	•Speech Synthesis •Corpora & Database •Machine Translation •language – Processing

Speech /NLP Activities in different Institutions (Cont….)

CDAC Kolkata (Bengali, Assamese, Manipuri)	•Speech Synthesis •Speech Corpora •Speech Recognition
CDAC Trivendrum (Malayalam, Tamil and Telugu)	•Speech Corpora •Speech Synthesis •Machine Translation
CFSL Chandigarh & CAIR, Bangalore (Hindi, Punjabi and other South Indian languages)	•Speaker Identification •Verification for Forensic Applications
AU-KBC Research Centre, Chennai (Tamil, Hindi)	•Machine Translation •Speech Recognition
Jadavpur University, Kolkata (Bengali)	•Machine Translation
Utkal University, Bhubneswar (Oriya)	•Speech Synthesis / databases •Machine Translation
ICS Hyderabad (Telugu, Hindi)	•Speech Synthesis
Lancaster University, UK, and CIIL, Mysore, India , (EMILLE Corpus), Most of the Indian Languages	•Text corpus – (Monolingual, Parallel) •Speech corpus
Prologix (Hindi)	•Speech Synthesis
Bhrigus Software (Hindi, Telugu etc.)	•Speech Recognition •Speech Synthesis
Lattice Bridge (Tamil)	•Speech Recognition
IBM India (Hindi)	•Speech Synthesis •Speech Recognition
HP Labs (Hindi, Indian English and other Indian languages)	•Speech Recognition •Speaker Identification
Webel Media (Hindi, Bengali)	•Speech Synthesis

Corpora Developments in Indian Languages

- **Text Corpora**
  - Kolhapur Corpus of Indian English (KCIE), Shivaji University, Kolhapur in 1988,
    - one million words of Indian English - for a comparative study among the American, the British, and the Indian English
  - TDIL programme, of DoE, Govt. of India initiated for development of machine-readable corpora of nearly 10 million words for all Indian national languages

Part	Language	Agency	Started	Closed	Word
I	English, Hindi, Punjabi	IIT,New Delhi	1991	1994	3 million
II	Telugu, Kannada, Tamil, Malayalam	CIIL, Mysore	1991	1994	Do
III	Marathi, Gujarati	DC, Pune	1991	1994	Do
IV	Oriya, Bangla, Assamese	IIALS, Bhubaneswar	1991	1994	Do
V	Sanskrit	SSU, Varanasi	1991	1994	Do
VI	Urdu, Sindhi, Kashmiri	AMU, Aligarh	1992	1994	Do

Corpora Developments in Indian Languages: Textual corpora

Central Institute of Indian Languages, Mysore	• data of 118 speech varieties, • materials such as grammar, dictionary, phonetic reader, report on dialect survey etc, • 3 million words each in Major Indian languages, • Anukriti - a data base on translation and translation studies, • LISIndia - a data base on Indian languages, creation of multilingual multidirectional dictionaries, • Indian classics translation - Katha-Bharati, • Digitization of library resources - Bhasha-Bharati • Development of machine-readable corpora of nearly 10 million words for all Indian national languages
C-DAC Noida	• GyanNidhi parallel textual corpus, • Aligned to paragraph level • 1 million pages of digitized parallel data from various sources in Unicode format
EMILLE project (Enabling Minority Language Engineering)	• written and spoken data for fourteen South Asian languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telugu and Urdu, • monolingual corpora approx 96,157,000 words • parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu.

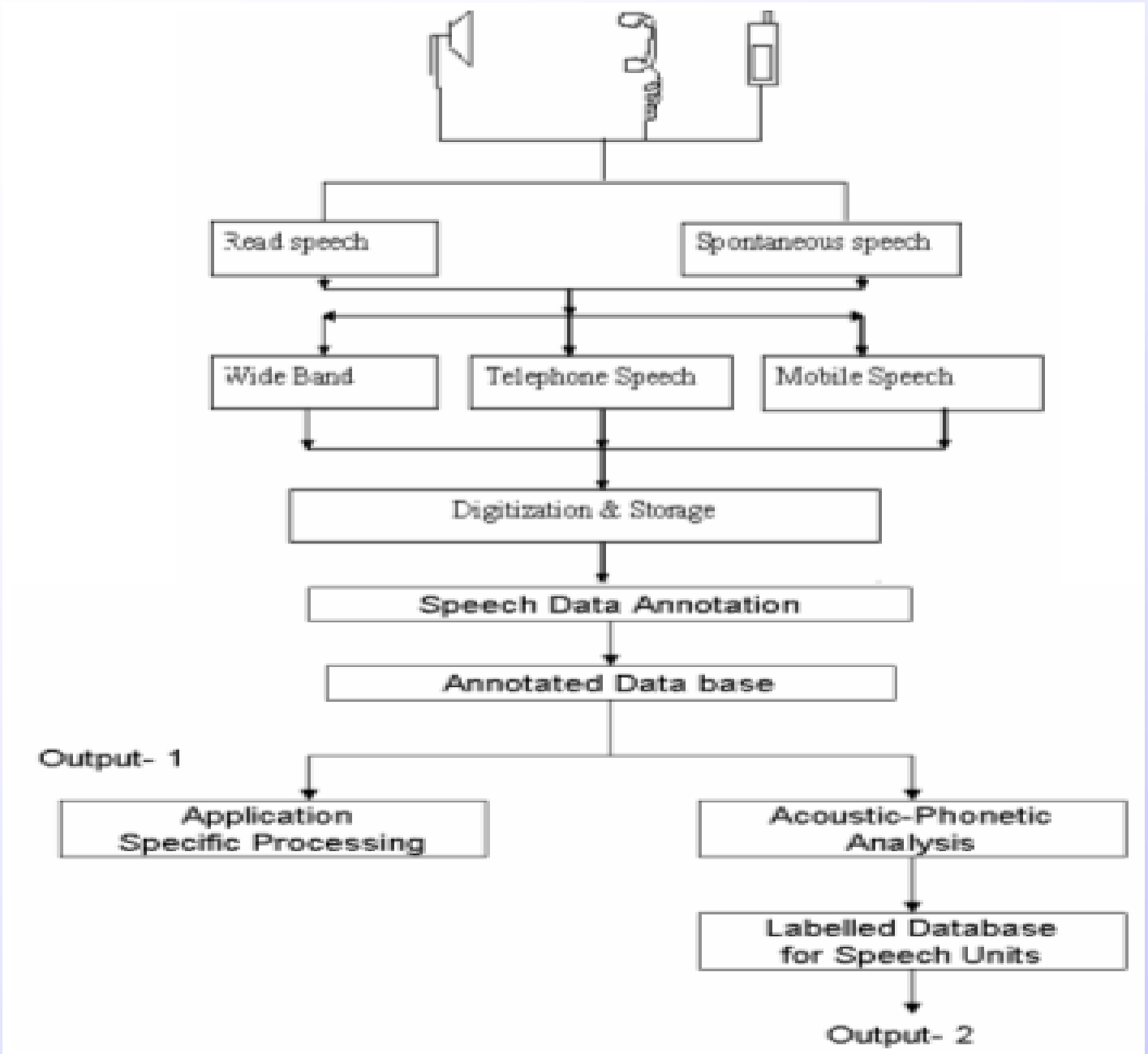
Corpora Developments in Indian Languages: Textual corpora

Mahatma Gandhi International Hindi University	• Started a project on databases and dialect mapping of Hindi called 'Hindi Samgraha,' • Digital warehouse of audio and visual data, • Interactive linguistic atlas, • Multi-dialectal lexicon and search mechanism for Hindi
Other major efforts	• The IIIT, Hyderabad lexical resources project to collect multilingual lexical data from Indian languages. • Tagging of MIT Bengali corpus at Indian Statistical Institute, Kolkata. • Anna University, Chennai has a text data on the specific topics on meditation and health of size 52,000 and 40,000

Speech Corpora Development Methodology at **CDAC,Noida**

- Four major steps
  - Selection of Textual content
  - Recording of Textual content
  - Annotation of Speech signal
  - Structural Storage of Corpora

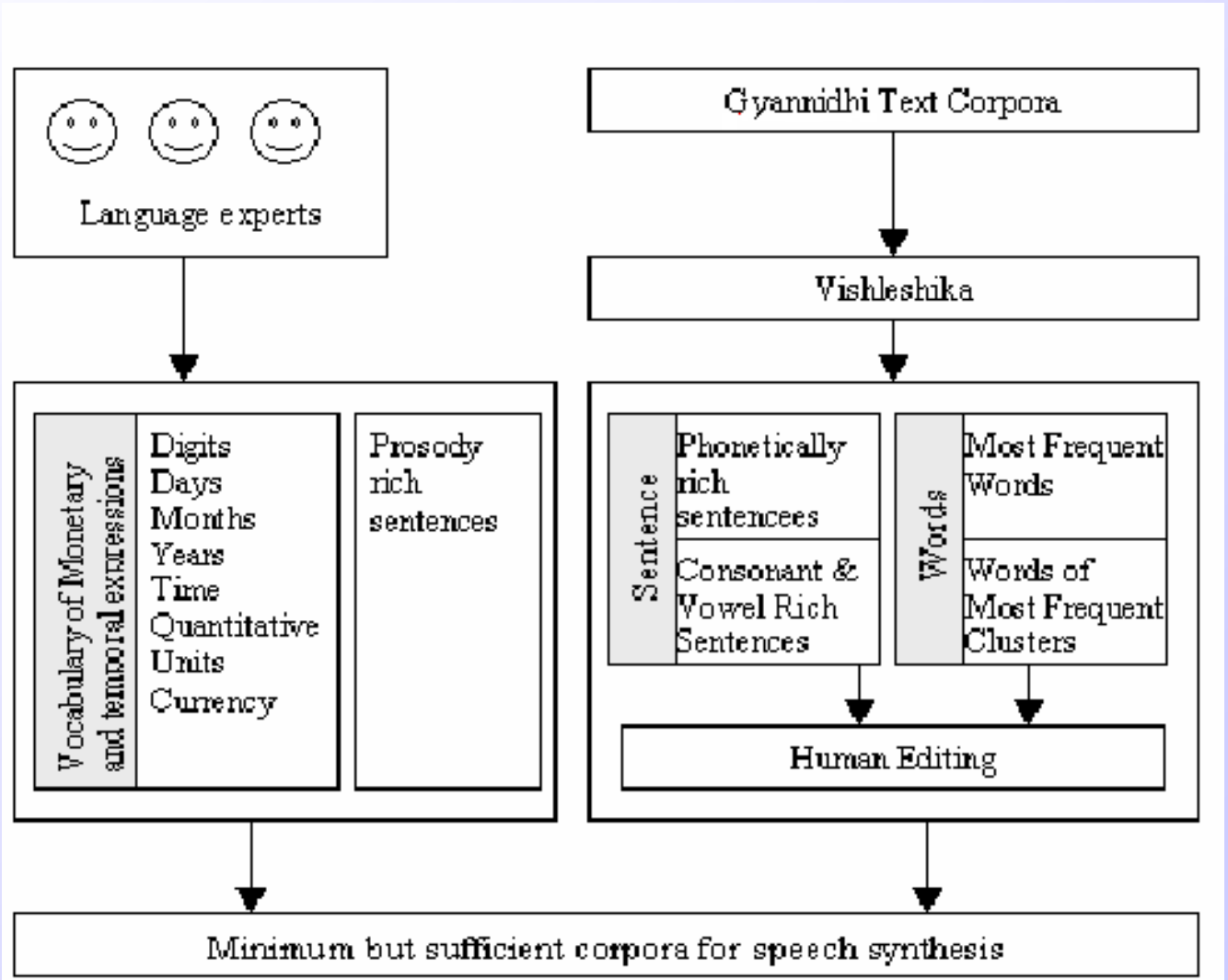
Speech recording Options



Corpora Development Methodology

- **Selection of Textual content**
- **Recording of Textual content**
- **Annotation of Speech signal**
- **Structural Storage of Corpora**

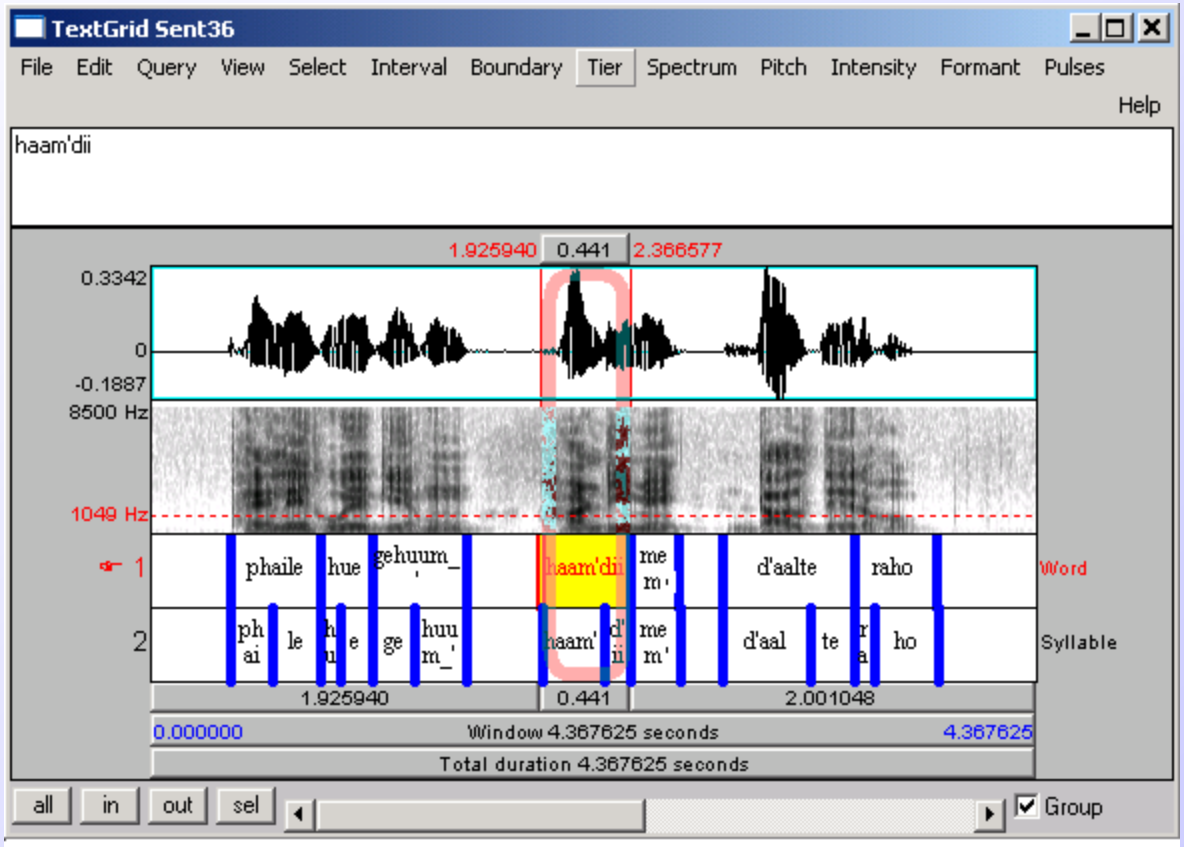
- Selection of Textual content



Labeling of sentence using Praat

फैले हुए गेहूँ हांडी में डालते रहो।

phaile hue gehum_' haamd'ii mem' d'aalte raho.



- Selection of Textual content
- Recording of Textual content
- Annotation of Speech signal
- Structural Storage of Corpora
  - META_DATA
  - TEXT_DATA
  - RAW_SPEECH
  - ANNOT_SPEECH
  - SPEECH_DATABASE
  - TOOLS

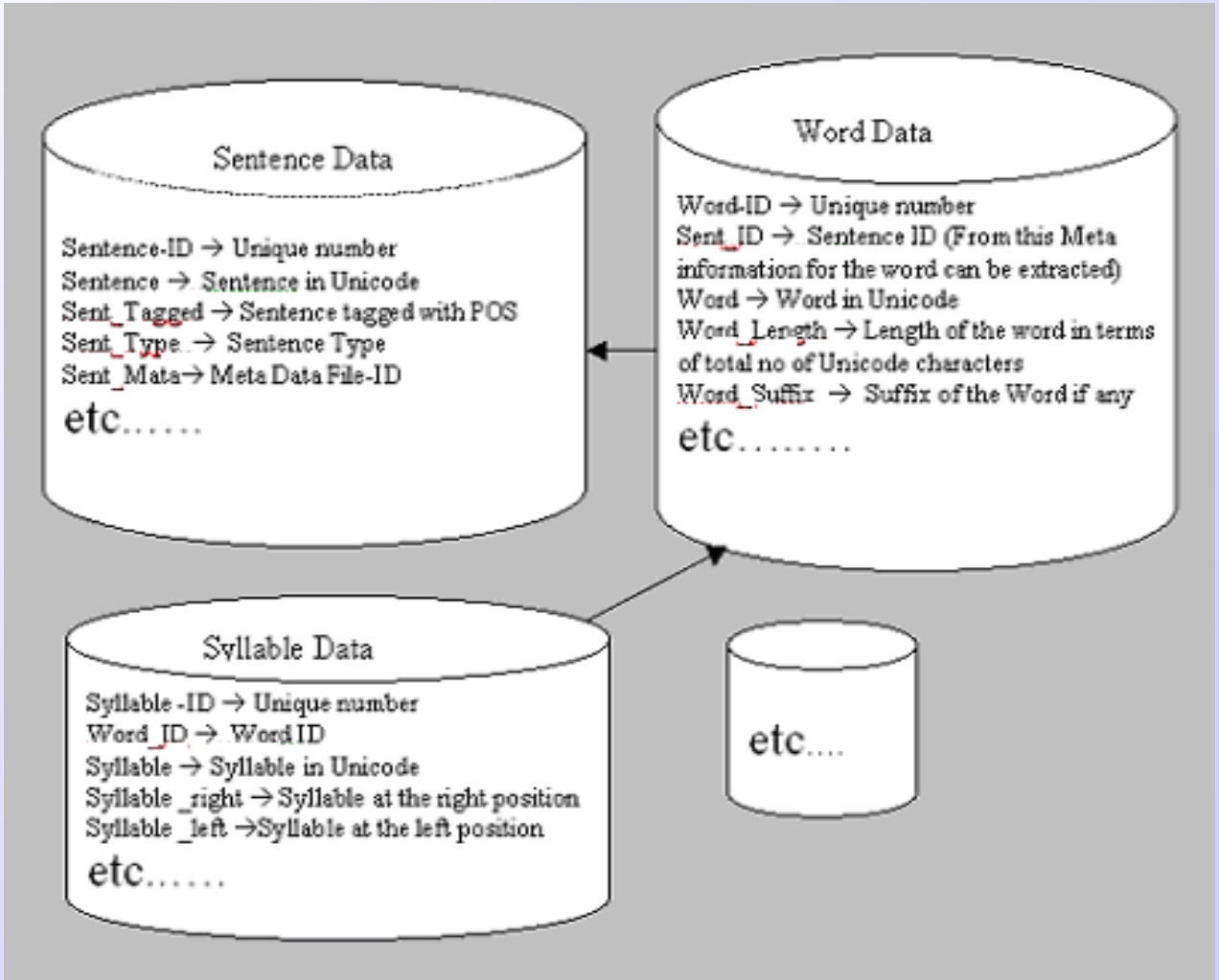
Corpora Development Methodology

- **Structural (Directory) Storage of Corpora**
  - META_DATA
  - TEXT_DATA
  - RAW_SPEECH
  - ANNOT_SPEECH
  - SPEECH_DATABASE
  - TOOLS

Corpora Development Methodology

- Selection of Textual content
- Recording of Textual content
- Annotation of Speech signal
- Structural Storage of Corpora
  - META_DATA
  - TEXT_DATA
  - RAW_SPEECH
  - ANNOT_SPEECH
  - SPEECH_DATABASE
  - TOOLS

- SPEECH_DATABASE



Speech Corpora for IL (Hindi, Punjabi & Marathi)

Speech corpora for Indian Languages (Hindi, Punjabi & Marathi)
Multi form phonetic data units

- Syllable,
- Most frequent words,
- Most frequent conjunct words,
- Vocabulary of digits, time, day, months, year, units
- Sentences of digits, time, day, months, year, units
- Phonetically Rich sentences
- Prosody Rich Sentences
- Domain Specific Text
- News Text
- Recording in noise free and echo cancelled studio conditions
- Recording by professional speakers (Male & Female) to maintain constant pitch & prevent stress phenomenon.
- Speech samples recorded at a sampling rate of 44.1khz (16 bit) in stereo mode
- Annotation of Speech units in a hierarchical manner, comprising of sentence, word, syllable
- Structural Storage of Corpora for ease in accessing
- Meta data for Speaker profile & Recording information
- User friendly interface for Speech Corpora view

Speech Corpora-DRDO

Speech corpora for IL (Manipuri, Assamese, Bengali)
Multi form phonetic data units

- Syllable,
- Most frequent words,
- Most frequent conjunct words,
- Vocabulary of digits, time, day, months, year, units
- Sentences of digits, time, day, months, year, units
- Phonetically Rich sentences
- Prosody Rich Sentences
- Domain Specific Text
- News Text
- Recording in noise free and echo cancelled studio conditions
- Recording by professional speakers (Male & Female) to maintain constant pitch and prevent stress phenomenon.
- Speech samples recorded at a sampling rate of 44.1khz (16 bit) in stereo mode
- Annotation of Speech units in a hierarchical manner, comprising of sentence,word, syllable.
- Structural Storage of Corpora for ease in accessing
- Meta data for Speaker profile & Recording information
- User friendly interface for Speech Corpora view

Hindi Speech Corpora: IN COLLABORATION WITH ELDA France

Speech of over 2000 speakers from different demographic profiles (age & sex), environments and dialects has been recorded over mobile (GSM / CDMA networks).

The speech data is annotated and a lexicon has been developed.

A wide range of utterances from isolated words, digit sequences, phonetically rich words and sentences to spontaneous responses are being recorded.

The speech database consists of:

- coverage of various dialectal variations in ratio of the populations speaking those dialects
- coverage of phonetically rich words and sentences
- coverage of speaking styles (commands, carefully pronounced and spontaneous speech)
- coverage of environmental influences (through mobile in various environments)

Corpus Design

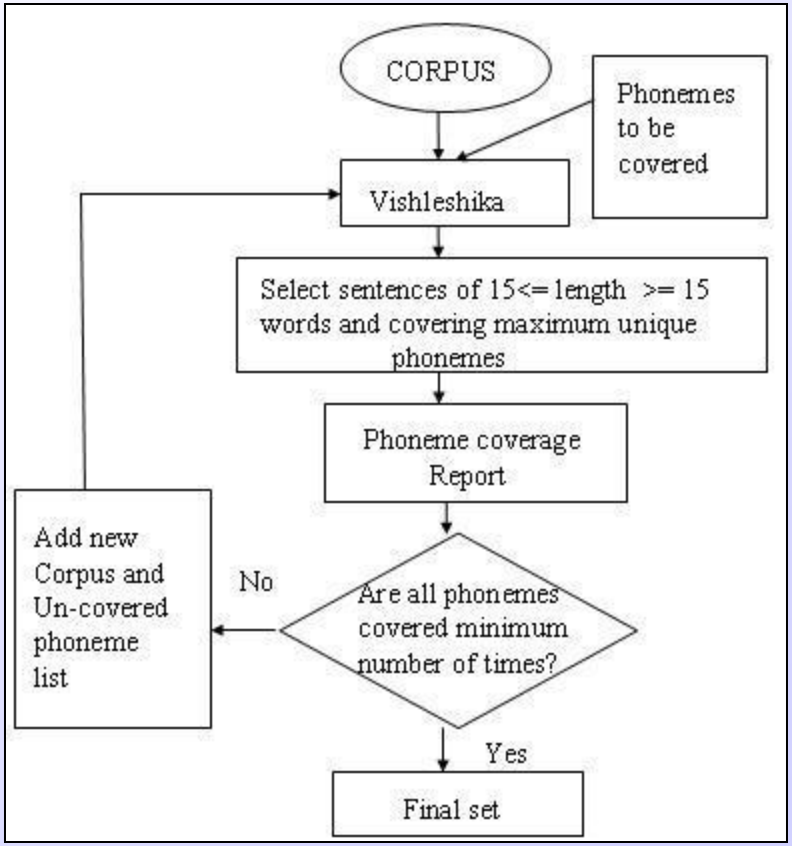
Based upon the specifications outlined in the LILA project, the corpus for Hindi was designed.

- In total there are 58 prompt items uttered by each speaker
- The vocabulary of the database contains digits and numbers (isolated digits, telephone numbers, PIN codes, credit card numbers, natural numbers, local currency), date and time expressions (months, days, holidays, time), directory names (city, company, forename and surname), phonetically rich material (words and sentences) as well as yes/no questions, spelt items and some control questions to keep track of the recordings.
- An additional item called silence word consists in recording 10 seconds of background noise without any speech.
- In the final corpus each speaker says four 'phonetically rich words' and 13 ‘phonetically rich sentences’. There are a total of 3204 different words and 7288 different sentences. Each word is repeated maximum five times and each sentence is repeated maximum ten times. Words were selected to have a good coverage of phonemes while sentences were selected to also have a good coverage of diphones.
- All sentences were chosen to have between 5 and 15 words and were individually checked to ensure that they are correct grammatically and in spelling, and that there is nothing potentially offensive or inappropriate in their content. Each speaker pronounces every phoneme at least once.

Corpus contents	Number of items
application keywords/key phrases	6
control questions	7
dates	3
digit/number strings	7
isolated digit items	3
money amount	1
natural number	1
phonetically rich sentences	13
phonetically rich words	4
proper names	5
spellings	3
times	2
word spotting phrases	1
yes/no questions	2
Total Prompts	58

Textual content creation mechanism

The corpus for selection of sentences and words were taken from news papers, Gyan Nidhi Parallel corpus and dictionaries. The database is fetched from corpus using a Statistical Analysis tool ‘Vishleshika’.



Specifications and distribution of speakers

For the purposes of this study, ‘Indian Hindi’ covered only persons who speak Hindi as a first language.

The 2000 speakers were divided into different demographic criteria (age, gender, network, environment and dialect regions) .

The database comprises 50% male and 50% female speakers, with a maximum deviation of 5%. Also, a minimum was imposed for the different age groups below.

Age groups

Age	Min.	Target
16-30	400 (20%)	654
31-45	400 (20%)	654
46-60	300 (15%)	492

Hindi is spoken throughout the Northern India. However only the states where Hindi is spoken in majority are being recorded.

The 18 selected dialects have been divided into 5 groups which represent mainly Western Hindi, Central & Eastern Hindi, Rajasthani, Bihari and Pahari.

The number of speakers of each dialect group is in proportion to the number of speakers in that region.

Group	Dialects(18)	Major States (9)	#speakers (2000)
Group 1	Khariboli, Brajbhasa, Hariyanvi, Bundeli	Delhi, Uttar Pradesh (west), Haryana,	1420
Group 2	Awadhi, Bagheli, Chattisgarhi	Uttar Pradesh (east), Chhattisgarh, Madhya Pradesh (east)	280
Group 3	Rajasthani, Marwari, Malvi, Mewari	Rajasthan	100
Group 4	Bhojpuri, Magahi, Maithili	Bihar, Uttar Pradesh (Cities on Bihar boarder)	100
Group 5	Pahari, Garhwali, Kumaoni, Himachali	Uttranchal, Himachal Pradesh	100

Recordings

The speech signal is recorded from the mobile telephone network (GSM, CDMA) via an ISDN line connection.

Recordings are being stored on 3 servers. The signals are stored directly in the digital format using A-law coding, with a sampling rate of 8 kHz, 8-bit quantization. A description of the sample rate, the quantization, and byte order used is stored in the label file of each utterance.

The following 5 acoustic conditions have been chosen as representative of a mobile user's environment:

- Passenger in moving vehicle, such as car, railway, bus, etc. (background traffic “emission noise”)
- Public place, such as bar, restaurant, etc. (background talking)
- Stationary pedestrian by road side (background traffic “emission noise”)
- Quiet location, such as home, office.
- Passenger in moving car using a hands-free car kit.

Transcriptions

For each signal file there is a corresponding label file in SAM label format to keep signal separate from annotation data, and it is extensible.

The SAM files also contained prompted text before transcription to help the transcribers. Thus, if the speaker pronounces exactly what was said, the transcriber needs only to confirm that it is correct and continues with the next transcription file. If not, changes are made by the transcriber to reflect what was said by the speaker and adds mark up from a minimal tag set.

The character set used for Hindi transcriptions is the Devanagari script and stored in UTF-8. The transcription is fully orthographic and includes a few details that represent audible acoustic events (speech and non-speech) present in the corresponding waveform files.

A set of markers for noises (non-speech items) and deviations like mispronunciations and recording truncations are used. Distortion markers include channel distortion, truncated waveforms, etc. and combinations of markers are used with pre-defined priorities.

Transcriptions, contd…

The tool used in the project is WebTranscribe, University of Munich and that was adapted to handle the Devanagari script in UTF-8.

This works in a distributed framework where the transcription data is stored on a server by means of a SQL database and can be easily accessed through a web interface by several transcribers.

To assure the quality of the transcriptions a number of procedures were established:

- transcribers went through a training
- guidelines were set up to harmonize the transcriptions and act as a reference
- annotators consult the transcription supervisor when in doubt.
- a second pass is done by another transcriber to cross check the data.
- a reference dictionary was chosen to align with standard spellings.

The transcriptions also included a romanized text version. Thus a romanization scheme had to be found. Existing INSROT scheme was modified to assure a one-to-one mapping.

Hindi phonetics and lexicon

The development of the database includes also a phonetic lexicon in SAMPA notation.

As no SAMPA notation existed for Hindi, a phonetic scheme for Hindi using SAMPA was drawn up in cooperation with the LILA consortium.

For each word in the database there is an entry in the lexicon together with the frequency for that word, romanized word form and the phonetic description in SAMPA.

Validation

Validation against specifications is being carried out by an independent validation center: SPEX, Netherlands.

The validation proceeds in three steps:

- Validation of prompt sheets in order to check the corpus before the recordings begin and to make sure it corresponds to the specifications.
- Pre-validation of a small database of 10 speakers. The objective of this stage is to detect serious design errors before the actual recordings start.
- Validation of complete database. The database is checked against the specifications and a validation report is generated.

Experiences and Current status

The recording supervisors have to remain attentive during the whole process of recording to ensure that the speaker do not take a very casual approach, and do the recordings completely and in a desired manner.

There was problem of echo sound in rainy season and hilly regions while recording in home / office environments.

There were the cases where beeps were getting recorded due to network problem and feedback. At some occasions the recording had to be repeated due to network failure while recording. This happened in the cases of moving environment, when the speaker crosses the cell boundary and enters another cell.

There were also cases where the recordings were saturated due to various reasons and most important the one were some speakers were speaking either very loudly or were keeping the mobile phone very close to their mouth.

All 2000 recordings have been completed including transcriptions. Care has been taken that no speaker is repeated in any of the given environments.

The database is collected according to the LILA specifications.

Conclusion

The final database consists of mobile phone recordings of 2000 native speakers of Hindi, recorded in five different environments (home/office, public place, street, moving vehicles and car kit recordings), of three age groups (16-30, 31-45, 46-60 years) and from five different dialectal regions.

Transcriptions are done in Devanagari script and include markers for speaker noise and non-speech events.

A lexicon with romanization, frequency and phonetics based upon SAMPA for each word in the database is also included.

The final product will be made available through the ELDA catalogue.

Text &Language Independent Speaker Identification for Forensic Applications

CFSL,Chandigarh&CAIR,Bangalore

VARIABLES

- Inter-speaker Variations-Repetitions, Health Condition, Age, Emotions,
- Contemporary/Non-Contemporary samples
- Same person Speaking different Languages
- Forced Variations-Disguise Conditions
- Environmental/Channel/Instrument Variations
- Type of Speech-Words, phrases, Sentences etc.
- (ROBUST SYSTEM REQUIRED)

Design of Data Base

Phase-I:

- Duration of speech for training: 15-20 Sec.
- Duration of speech for testing: 5 Sec.
- Type of samples : Isolated, Contextual and Spontaneous

- No. of languages : Ten (10)

(Hindi, English, Punjabi, Kashmiri, Urdu, Assamese, Bengali, Telugu, Tamil, Kannada)

- No. of speakers: Ten (10) in each language (Total: 100)

Design of Database…Contd

- (a) Multi-lingual (10 languages)-

- Hindi, Punjabi, Urdu, Bengali, Assamese,
- Telugu, Tamil, Kannada
- Kashmiri,
- Indian English

Design of Database Contd….

- Multi Channel— 10 different channels
- Three hand held microphones-Dynamic, Condenser, Computer desk top
- One Telephone Hand set (PSTN)
- One Telephone Handset (CDMA)
- One Mobile phone handset (GSM)
- One headset output
- Three different Tape recorders-

Speaking Conditions

- Each Speaker in Three different Languages—Mother tongue, Hindi and Indian English
- Each speaker speaking in two different sessions-Time difference min. of six months
- Two recordings in each session-15 seconds,5 seconds (non-repetitive phrases)
- Two minutes of effective speech from each speaker.
- Isolated words, Contextual sentences

Recording/Digitization

- No. Of Speakers --- 100 (10 native speakers of each language)-I phase
- Damped and Noisy conditions
- Disguise conditions (mimicry, pencil etc.)
- PA-Pre Amplifier (TASCAM System,DM-3200,Digital Mixing console)
- 96000Hz./48000Hz.
- 16/24 bits/sample.
- .wav files

Contd..

- Non contemporary:

  - ***Samples recorded in different interval of time (Time gap on minimum 6 months and maximum 1year)***

  - ***Samples recorded with different recording devices e.g training samples are recorded one device and testing samples are from different device***

- No. of modes : Direct (six microphones), telephone, mobile phone, mobile with noisy (Car , Traffic light) and with three recorders

Continue….

- No. of utterance :Three (Two at same time & one after six months)
- Duration of utterance : 2 minutes
- Type of samples : Isolated, Contextual and Spontaneous
- Total No. of the samples : No. of speakers X No. of modes X type X No. of languages X No. of utterance

(100 X 12 X 3 X3 X 3 = 32400)

Phase-II

System should be designed/developed on the basis of disguised mode of speaking

Modes of disguised recordings:

- Handkerchief in front of the mouth
- Chewing of betel leaves
- Cigarette or pencil in the mouth
- Closing of nose
- Artificial disguise

NEED TO BENCH MARK

- There is a great need to develop appropriate speech database in different conditions (different languages and channels etc.) and to bench mark and justify the utility of speaker identification system for Forensic Applications.

Contd…

- CFSL, Chandigarh developed the database consisting of 100 speakers in ten different languages as well as in eleven devices.
- Prototype system developed by CAIR, Bangalore for language independent speaker identification is in testing stage

Results of Prototype testing

- Training environment:D04 English Contextual
  (size of Training file: 50-60 Sec.)
  (size of Testing file: 20 Sec.)

- Testing environment : Number of speakers
  - Hindi and English:20 each
  - Punjabi:10

Test Results

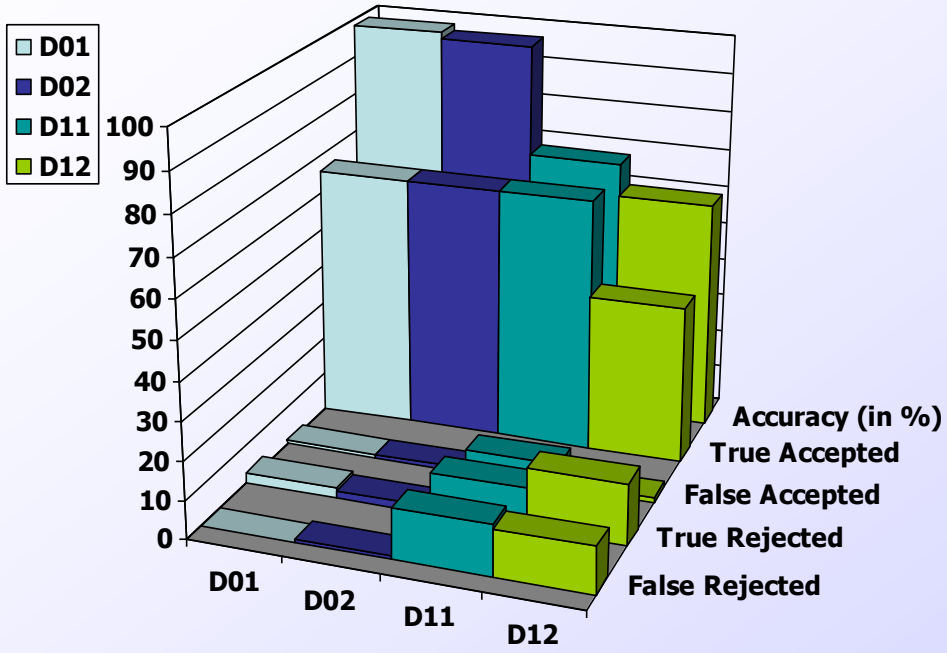
Mode	Pos	Exp-1HC		Exp-2 EC		Exp-3 HI		Exp-4 EI		Exp-5 S	
		Spr.	%	Spr.	%	Spr.	%	Spr	%	Spr	%
D01	I (II)	19 (1)	95	19 (1)	95	14 (1)	70	17 (2)	85	13 (4)	65
D02	I (II)	18 (2)	90	18 (1)	90	14 (2)	70	18 (1)	90	16 (1)	80
D03	I (II)	14 (1)	70	14	70	10 (1)	50	14 (2)	70	11 (3)	55
D04	I (II)	18 (1)	90	19	95	16 (2)	80	18 (1)	90	18 (1)	90
D05	I (II)	17 (2)	85	19	95	15 (2)	75	17 (2)	85	19 (1)	95

Test Results

Contd….

Mode	Pos	Exp-1HC		Exp-2 EC		Exp-3 HI		Exp-4 EI		Exp-5 S	
		Spr.	%	Spr.	%	Spr.	%	Spr.	%	Spr	%
D06	I (II)	15 (3)	75	19	95	12 (4)	60	17 (2)	85	13 (3)	65
D012	I (II)	10 (1)	50	15 (4)	75	05 (4)	25	06 (2)	30	09 (3)	45
D013	I (II)	13 (5)	65	15 (3)	75	08 (3)	40	08 (5)	40	11 (2)	55

Testing results on Punjabi :
Windows Version



Train: D04 (HC)

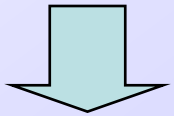
Test: D01,D02,D11 &D12 (EC)

No of speaker: 10

Device	Accuracy	True Accepted	False Accepted	True Rejected	False Rejected
D01	98.57%	66	1	3	0
D02	97.14%	66	1	2	1
D11	68.18%	66	06	11	13
D12	60.04%	41	1	16	12

Objectives

- A-STAR
  - accelerates the development of large scale spoken language corpora in the Asia.
  - advances related fundamental technologies such as
    - multi-lingual speech translation
    - multi-lingual speech transcription
    - multi- lingual information retrieval



Speech Translation

Goal

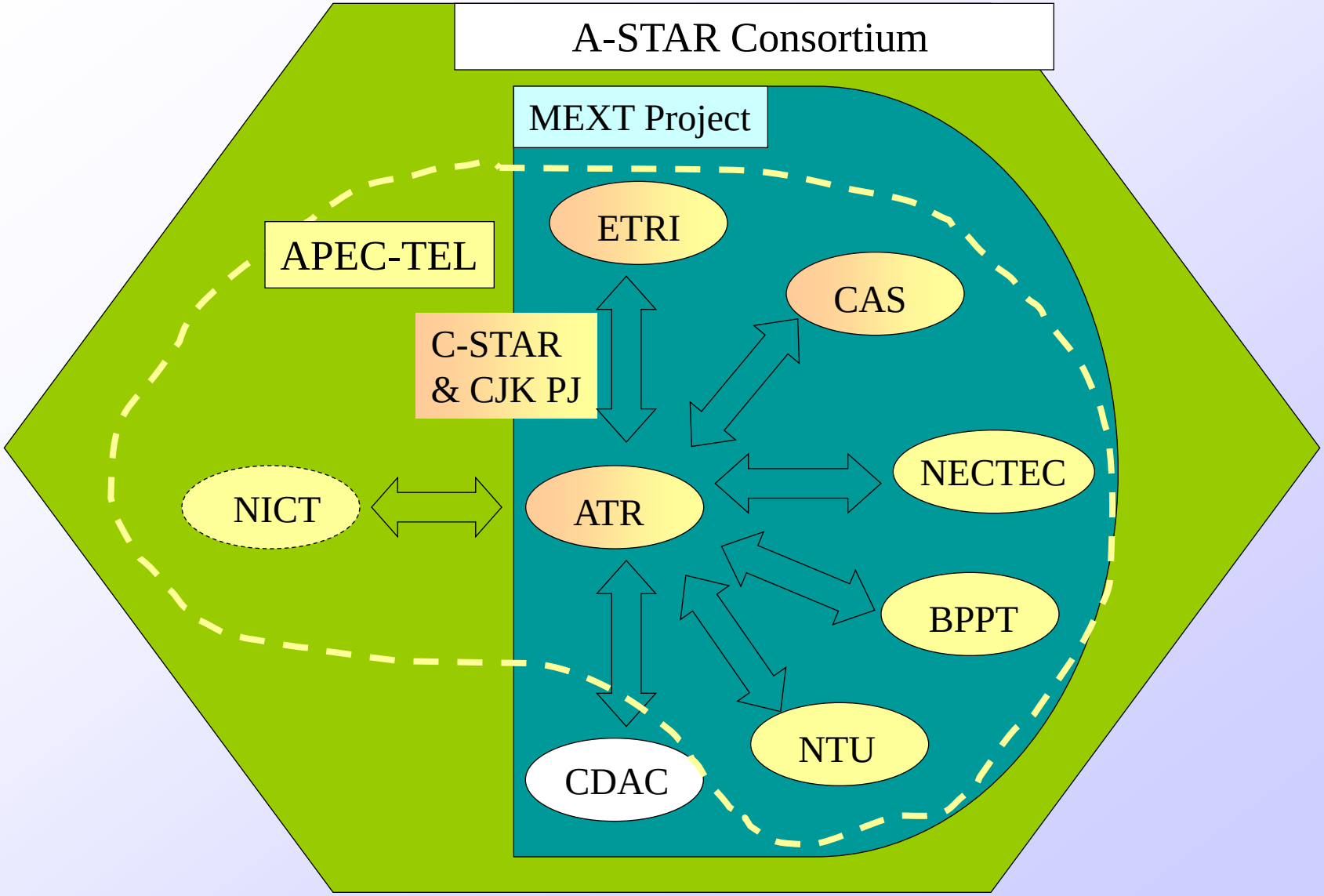
- This project aims:
  - Establishment of an **international research collaboration group**
  - Building **large scale speech and language corpora and technologies**
  - Initiate **speech translation trial service** in Asia

- Target languages:
  ATR (Japan, coordinator), NLPR(China), ETRI(Korea), BPPT(Indonesia), NECTEC(Thailand), CDAC(India) and National Taiwan Univ. (Chinese Taipei) will start the investigation, and seek and choose their partners for the other languages in Asia.

- A-STAR has partly been approved by MEXT and APEC-TEL

A-STAR members

Region	Institutions
Japan	Spoken Language Communication Res.Labs., ATR (Coordinator), Dr. Satoshi Nakamura
China	NLPR, Institute of Automation, Chinese Academy of Sciences, CASIA, Prof. Bo Xu
Chinese Taipei	National Taiwan University, Prof. Linshan Lee
Korea	Electronics and Telecommunications Research Institute, ETRI, Dr. Young Jik Lee
Indonesia	Agency for the assessment and application of technology, BPPT, Dr. Hammam Riza
Thailand	National Electronics and Computer Technology Center, NECTEC, Dr. Chai Wutiwiwatchai
India	Centre for Development of Advanced Computing, Prof. Shyam Sunder Agrawal

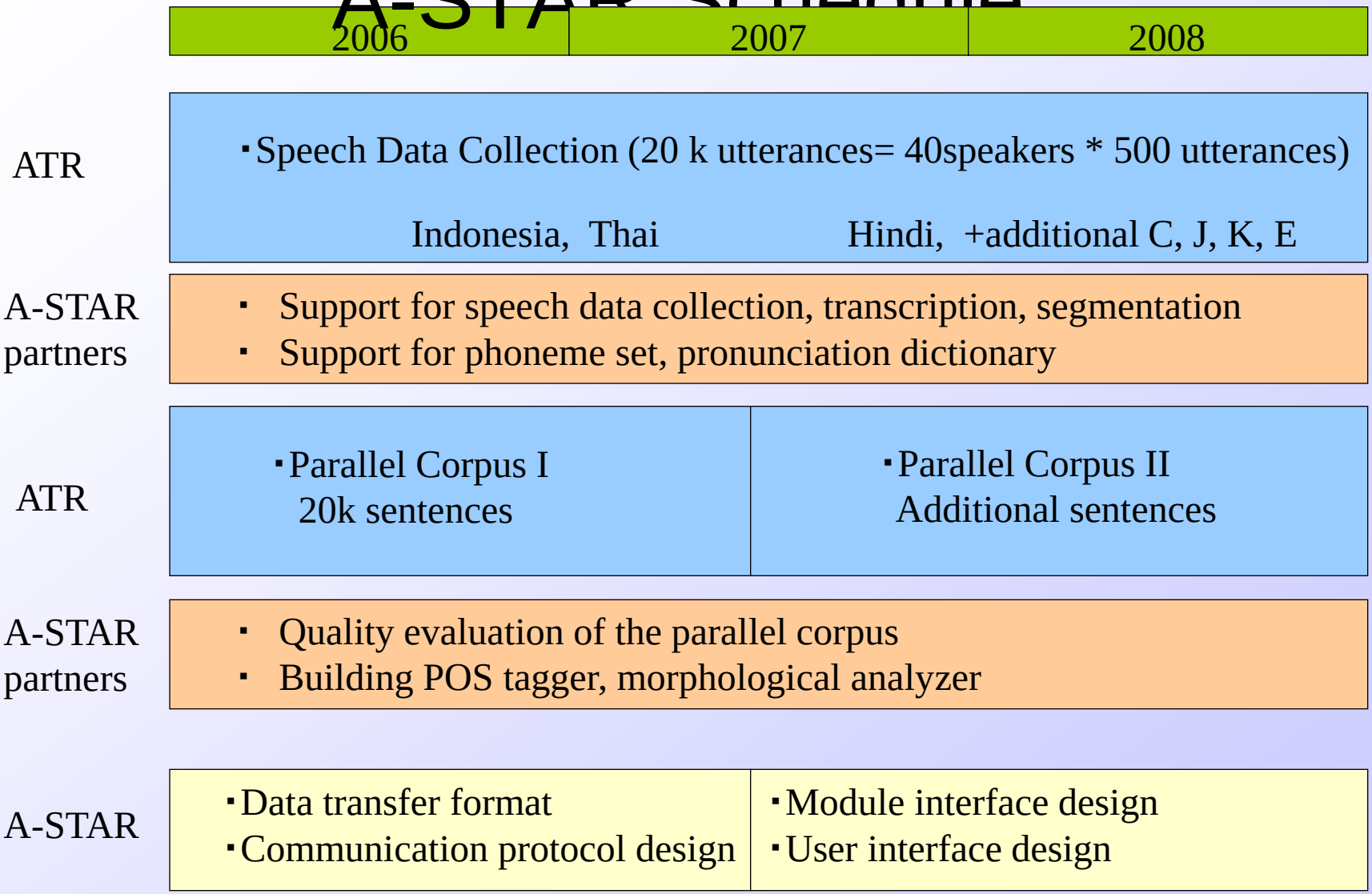


A-STAR Consortium covers all of the activities !

What to do?

- Corpora
  - Standardize parallel corpora format
  - Standardize communication protocol
  - Collect fundamental parallel corpora in Asian languages
- Format of linguistic tag information
  - Morpheme, pronunciation, intonation
  - Entries of the dictionary
- Communication protocols of modules
  - Interface formats among speech recognition, speech synthesis, and language translation necessary for speech translation
  - API formats of speech translation and modules for developers

A-STAR Schedule



Corpora Developments in Indian Languages: Speech corpora

CEERI & TIFR	• 207 words, 50 speakers (two utterances), for development of voice operated Railway Reservation Enquiry Systems . • 1000 phonetically rich sentences, 100 speakers, phonetically segmented and labelled, for development of phoneme based speech recognition systems. • 1000 most frequently used words of Hindi, 50 speakers, a database of English and Hindi connected digits etc.
C-DAC Noida	• Hindi, Marathi and Punjabi (jointly with CSIO). • Phonetically rich sentences and most frequent word set (10,000) from GyanNidhi corpus using 'Vishleshika' – A Statistical Text Analyzer tool. • Prosodically representative sentences set (1000 sentences), 16-bit PCM mono 44.1KHz for development of speech synthesis systems.
C-DAC Kolkata	Bengali, Assamese and Manipuri languages, Phonetically balanced word-set, prosodically representative sentences set (850 sentences) Most frequently spoken word set (11,000 words), 16-bit PCM mono 22050Hz for development of speech synthesis system and also ASR systems.

Corpora Developments in Indian Languages: Speech corpora

HP Labs India	• Hindi & Indian English databases for TTS, • Assamese and Indian English database for ASR • Currently, collection is underway for Marathi, Tamil and Telugu in collaboration with IIIT, Hyderabad.
Utkal University, Bhubneswar	• 50 words paragraph recording by different speakers • Recording in noise free environment with five times recordings • Most clear recordings are chosen out of those.
IIT Kharagpur	• 180 speakers (60 for each for Marathi, Hindi and Urdu) • 22,050 Hz, 1 channel, 16 bit resolution and 10 repetitions. • Various environmental conditions road/home/office/slums/ college/train/hills/valleys/remote villages/research labs/farms • For Text-independent speaker identification in ASR
IIT Madras	• Hindi and Telugu, Doordarshan news bulletin • Single speaker, segmented at phoneme level • For Telugu, 14 minutes and 156 read sentences. • For Hindi 12 minutes and 121 read sentences
IIT Bombay	• A prototype TTS named Vani based on concatenation synthesis. • Fract-phoneme as the basic unit

Corpora Developments in Indian Languages: Speech corpora

CFSL, Chandigarh	• Speaker Identification Database (SPID) for English and Hindi. • Isolated and contextual speech samples used in conversation • Recording in 10 different modes of normal and disguised speaking • For forensic applications.
Utkal University, Bhubneswar	• Multi-channel Isolated word database for Indian English 77 words • Spontaneous speech in local language. • For speech recognition algorithms for use in telephony system.
Bharati Vidyapeeth, Pune	• Approximately 5,000 words • A speech synthesis system for Marathi based on word concatenation approach • For reading Online daily ‘Sakal’
ICS, Hyderabad	• Telugu and Hindi words as Macro-syllabus. • For development of Interactive Voice Response Systems

Speech Synthesis

Institute	Language Cover	Synthesis Strategy	Unit/Database	Text/Speech Segment processing/ Tools	Prosody	Performance
CEERI, Delhi	Hindi Bengali (partly)	Formant (Klatt – type) Synthesis	Syllables & Phonemes (Parameter Data Base	Manual (Rules for smoothing) Parsing rules for Syllabification	Manual + Some rules	Copy Synthesis Excellent Unlimited TTS-Average
TIFR Mumbai	Hindi Bengali Marathi Indian English (Partly)	Format (Klatt – type) Synthesis	Phonemes and other units	Automatic parsing rules for phonemization, Rules for smoothing prosody	Prosody rules	Unlimited TTS – More than Average
IIIT (Hyd)	Hindi Telugu Other languages	Concatenative	Data base in required languages as per festival norms	For unit as per festival system As per requirements of Festival System	Prosody studies in required language done and implemented	Un-limited TTS- better than Average
IIT, Chennai	Hindi Tamil	Concatenative di-phone synthesis (1400 diphones)	Syllabus (Mainly)	Automatic segmentation using group delay functions for unit selection Festival System	Pitch tracks determined and implementation	Unlimited TTS-Average
CDAC, Pune	Hindi, Indian English	Concatenative	Phonemes, other units			
CDAC, Noida	Hindi	Concatenative	Multi – form units –Syllables, frequent words, phrases etc.	Parsing for syllables, Statistical processing of text for formation of phonetically rich sentences and other units (Vishleshika)	Study of intonation patterns,	Domain Specific- Excellent, Unlimited TTS-Average
CDAC Kolkata	Bengali	Concatenation	Phonemes & Sub – Phonemes (Size 1 MB)	Cool – edit Phonemic/ Segmentation	TDPSOLA /ESNOLA	Unlimited TTS-Average
Bhrigus Software Ltd. Hyd	Hindi, Telugu & Others	Concatenative	Phonemes, Using Festival requirements	Fest VOX tools Festival	Intonation using (CART for Prosody modeling)	Unlimited TTS-Average
Prologix Software, Lucknow	Hindi	Concatenative	Di-phone data base	Festival based- Fest VOX tools		Unlimited TTS-better than Average
Webel Mediatronics, Kolkata	Bengali, Hindi	Formant type	Phonemes (Parameters of phonemes)	Rules for concatenation and smoothing of parameters Text processing rules	Intonation rules being implemented.	Unlimited TTS.- Less than Average

Integration of OCR &TTS with Hindi Unicode Word Processor

- Unicode Word Processor named Swarnakriti, with basic features of Word Processor, like editing, printing, formatting & typing in InScript for Indian Languages features,
- Special features like Spellchecker for Hindi and English to Hindi Transliteration are embedded
- Embedded utilities like…
  - Calculator
  - Calendar
  - Various TTS from different developers have been tested, discussions with developers is under process, but yet not finalized.
  - Prototype TTS integration has been tested and demonstrated during ELITEX.

Speech Recognition

Institute / Type of system	Languages	Technology	Performance Usage
1 CEERI (i) IWRS (Limited Vocabulary) (ii) Corrected words / Isolated words	Language Independent Hindi	Hardwired -Filter Bank Micro-Proc. Based System Distance Comparison VQ / MFCC / FBMC / TDNN based	Wheel Chair Robotics Laboratory Testing / Comparison with German Language
2. IIT, Chennai Sub – words : Units	Tamil	Comparison of HMM based classified with support vector machine based classifiers	Laboratory experiment
3. IIT, Kanpur	Hindi	Modification of MFCC and comparison with Weighted Overlapped Segment Averaging HTK toolkit.	Laboratory experiment
4. IIT, New Delhi	Hindi	Language modeling for Speech Recognition, Using synthetically enhanced latent Symantec Analysis	Laboratory Expts.
5. TIFR, Mumbai (i) Isolated Words (Medium SizeVocab.) (ii)Continuous Speech	Hindi Hindi/Other Indian Lang.	Feature Based Hierarchical ASR HMM Based using Sphinx tools	Computer Tutor / Travel guide system Desktop Telephone Number Desktop / Telephone Based SRS
6. IBM (I) Continuous Speech Recognition	Hindi	Adaptation of IBM via voice HMM Based acoustic recognition using Trigram Language Model (Mapping to Hindi phonemes)	
7. CDAC, Kolkata Word / Subword units	Bengali	Lexically driven manner based Recognition	Laboratory Expts.
8.IISc.Bangalore		Inhomogeneous HMM	Laboratory Expts

Machine Translation

Institute	Technology/System name	Language Pair	Approach	Remarks
IIT Kanpur CDAC Noida	AngaBharti	English - Hindi	Pattern directed Rule Based with embedded Example base	Technology developed by IIT Kanpur. Technology is being adapted for other target Indian languages
CDAC Pune	Mantra	English-Hindi	Tree Adjoining Grammar based	For Government notifications
CDAC Mumbai	Matra	English	Frame based	For News stories
IIIT Hyderabad	Shakti	English –Hindi English-Telugu English-Marathi	Example Based	
IISc Bangalore, IIIT Hyderabad	Shiva	English to Hindi Hindi to English	EBMT paradigm	
IIIT Hyderabad	Anusaarka	Kannada-Hindi, Marathi-Hindi, Punjabi-Hindi, Telugu-Hindi, Bengali-Hindi	Example based	
IIT Mumbai		English,Hindi, Marathi	Interlingua (UNL) based	
CDAC Kolkata	AnglaBharti	English-Bengali		
AU-KBC Research Centre, Chennai		English-Tamil Tamil-Hindi	HMM based Statistical Method	English-Tamil for Traditional Indian Medicine domain
Jadavpur University, Kolkata		English-Bengali	Phrasal Example based	Research for News Headlines translation

Special Tools for Text / Speech Processing

- Vishleshika- statistical text processor
- Prabandhika- Corpus Manager (Corpus data in user defined domains)
- Lekhika- Indian Language Word Processor
- Shabdika- Dictionaries providing corresponding meaning of English in Hindi
- CLIR- Cross Lingual Information Retrieval
- Multi Lingual Crawler- Information Retrieval System
- Text summarisation
- Annotation of Text and Speech
- Spell Checkers
- Unicode conversion tool
- Large dictionaries
- Parallel Corpora (Gyan Nidhi)
- Tagging of Corpora

LEKHIKA- A PLATFORM INDEPENDENT WORD PROCESSOR

A word-processor with tools like

- Dictionaries
- Translation /transliteration,
- Powerful spell-checker in local languages
- Desktop utilities like calendar, calculator, Unit converters etc.
- ISCII to UNICODE converters

FEATURES

- Multi lingual document support
- Preserves Font, Style & Language etc.
- Multiple Document Interface
- Native system look and feel
- Print preserving their Font, Style & Language
- Multilingual Help
- Embedded Dictionary Spell Checker for local languages
- Translation of any word
- Translation/Transliteration Facility on the basis of
  - Any line
  - Any selected portion of text
  - Any Document
- Utilities like Calendar, Calculator

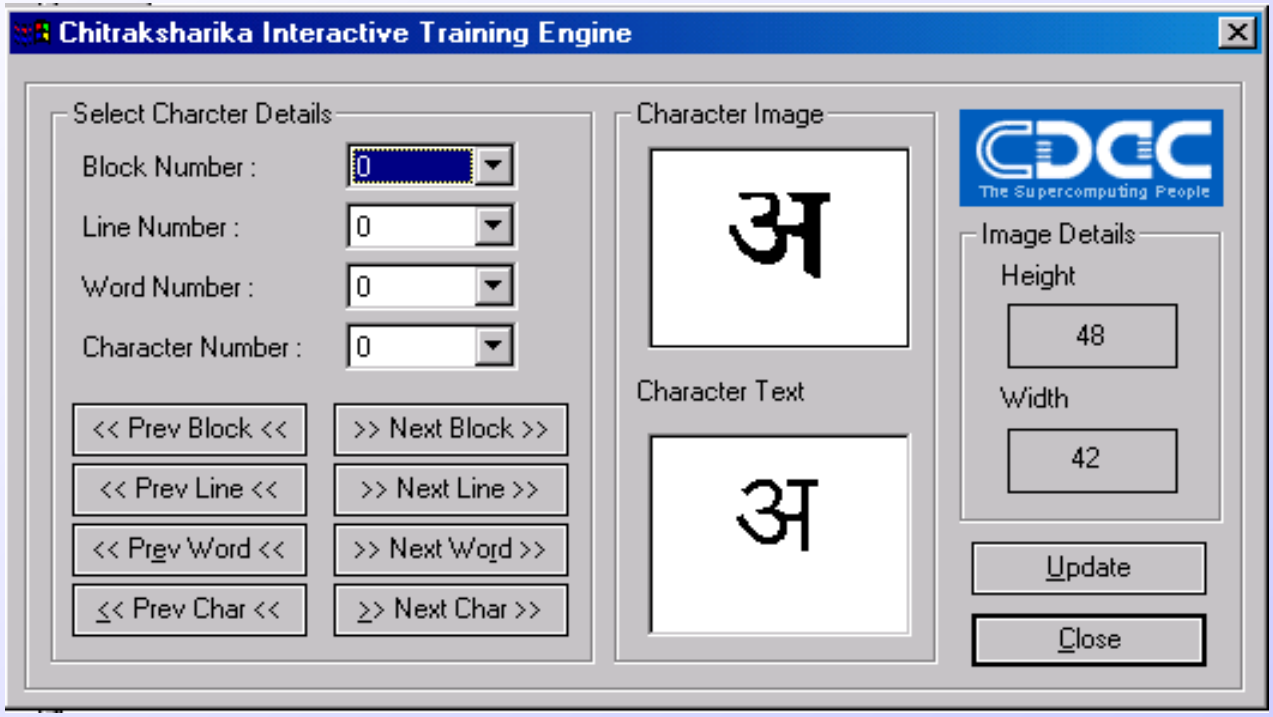
CHITRAKSHARIKA:OPTICAL CHARACTER RECOGNITION FOR DEVNAGARI

Features

- **Image Binarization**
- **noise cleaning**,
- **text block identification**,
- **skew correction**,
- **line and word detection**,
- **character segmentation**,
- **character recognition** and **error correction**
- **Training Engine**

Template Addition (Training Engine)

The main GUI for the Training Engine is shown below:



SHABDIKA

This is a package of various dictionaries providing the corresponding meaning Hindi of English term.

Features

-User Friendly GUI
-History of last used words.
-Categorized look up
-Related words storage
-Fast retrieval of information
- Authenticated Source of Information

Tagging of Hindi Corpora

•Corpus Collected from CIIL Mysore was proof read and corrected for mistakes

•Categorized Corpus in following categories

- Aesthetics
- Social Sciences
- Natural, Physics & Professional Sciences
- Commerce
- Official and Media Language
- Translated Material

•Tagger / Morphological Analyser provided by Anusaaraka Group, Morphological analyzer was Modified to get improved tagging

•Rules framed with help of KHS to improve tagged output

•Development of software utility for Romanization of tagged corpus

•GUI Development to view corpus with grammatical tags and information

•Tagged corpus uploaded on TDIL webserver and data in CD with user interface

•Hindi Corpora of about three million words has been developed on the basis of literature published in Hindi. It is a sort of General Corpora with a collection of texts of different types and is a source for studying various features of the language in general. This Corpus has been prepared on the basis of 76 subjects.

On-Line Hindi Vishwakosha (Hindi Encyclopaedia)

- A joint project of ***KHS, Agra (MHRD)*** and ***CDAC (DIT)*** for bringing out the Hindi Encyclopedia (published by ***Nagri Pracharini Sabha, Varanasi***) on the net in public domain
- User-friendly interactive GUI
- More than 15,000 topics
- Information arranged in Alphabetical as well as categorized form
- Search in Hindi within the site
- Facility to search in Hindi without having to key-in
- Site Contents are changed every time the site is loaded
- Site has been enriched with images where ever necessary
- Gist of the topics have been provided at front screen to help user in tracing the desired information
- “Do you knows” have been added to attract children and general surfers

Home Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

हिन्दी विश्वकोश

आक्सनार्ड

आक्सनार्ड नगर संयुक्त राज्य, अमरीका, के कैलिफोर्निया राज्यांतर्गत वेंट्यूरा जिले में, सेंटा बारबरा चैनल के तट के समीप, लास एंजिल्स नगर से पश्चिमोत्तर पश्चिम दिशा में 50 मील की दूरी पर स्थित है।

और जानें

अस्थिचिकित्सा

शल्यतंत्र का वह विभाग है, जिसमें अस्थि तथा संधियों के रोगों और विकृतियों या विरूपताओं की चिकित्सा का विचार किया जाता है।

और जानें

आजाद

द्र चंद्रशेखर आजाद।

और जानें

आस्ट्रिया

मध्य यूरोप के दक्षिणी पूर्वी भाग में एक छोटा गणतांत्रिक राज्य है।

और जानें

अल्बुला

स्विट्जरलैंड के ग्रिसन नामक पहाड़ी भाग का एक प्रसिद्ध गिरिपथ है।

और जानें

इंजन

उस यंत्र या मशीन को कहते हैं

- धर्म
- राजनीति
- कला एवं संस्कृति
- साहित्य
- विज्ञान
- क्रीड़ा
- इतिहास

- विज्ञान
- कला एवं संस्कृति
- धर्म
- व्यक्तित्व
- साहित्य
- भूगोल
- संस्था
- इतिहास
- नाथा
- रक्षा
- अर्थशास्त्र
- शिक्षा
- क्रीड़ा

Done

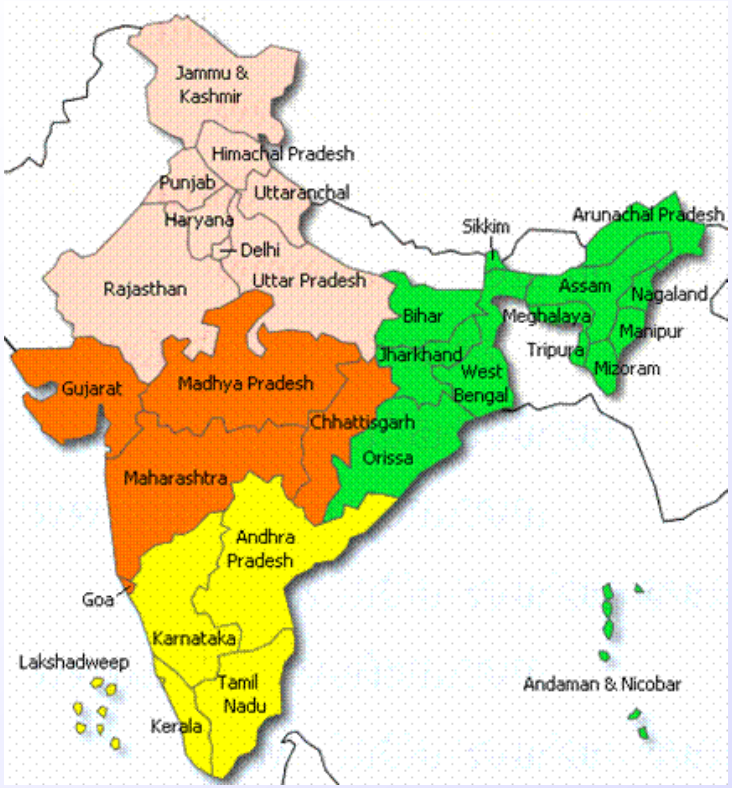
Local intranet

On-Line IT Terminology in Hindi

- A joint project of CSTT, New Delhi (MHRD) & CDAC, Noida (DIT) for bringing out the Information Technology Terminology in Hindi on the net in public domain
- User-friendly interactive GUI
- Collection of around 10,000 standardized terms with their Hindi equivalents
- Search facility within the site in English as well as Hindi
- Categorization of terms in various fields of Information Technology
- Displays Word of the Day for casual surfer
- Displays Random words for casual surfer
- Site is Bilingual i.e the content can be seen in Hindi as well as English as base language
- Facility to search in Hindi without typing
- Site comes with free font in public domain available for download
- Files available in categorical and alphabetically for downloading
- Site available on TDIL web server **www.tdil.gov.in**


Address http://yuvraj/itterminology/
Go
राष्ट्र की एकता को यदि बनाकर रखा जा सकता है तो उसका माध्यम हिंदी हो सकता है।
सुब्रहमण्य भारती
Hindi / English Search
खोजें
Search
आज का शब्द
यादृच्छिक शब्द
होम
अक्षरात्मक
श्रेणीगत
डाउन लोड
योगदान
सुझाव
हमें जानें
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
शब्द
हिन्दी रूप
श्रेणी
Daemon
डेमॉन
Operating System
Daignostic system
निदानी तंत्र
Graphical / Multimedia System
Daisy print wheel
डेज़ी मुद्रण चक्र
Storage Devices
Daisy wheel
डेज़ी चक्र
Storage Devices
Data acquisition computer
आंकड़ा अर्जन अभिकलित्र
Computer Architecture
Data administrator
आंकड़ा प्रशासक
DataBase System
Data allocation
आंकड़ा नियतन
Operating System
Data aquisition
आंकड़ा अर्जन
DataBase System
Data area
आंकड़ा क्षेत्र
DataBase System
Data array
आंकड़ा सरणी
DataBase System
Data atribute
आंकड़ा गुण
DataBase System
Data base
आंकड़ा आधार
DataBase System

Gyan Nidhi : Parallel Corpus



Language	States
Hindi	Uttar Pradesh, Delhi, Bihar, Rajasthan, Madhya Pradesh
Punjabi	Punjabi
Kashmiri	Jammu & Kashmir
Urdu	Jammu & Kashmir
Bengali	West Bengal
Oriya	Orissa
Assamese	Assam
Manipuri	Manipur
Nepali	Sikkim
Marathi	Maharashtra
Gujarati	Gujarat
Konkani	Goa
Kannada	Karnataka
Telugu	Andhra Pradesh
Tamil	Tamilnadu
Malayalam	Kerala
Sanskrit	Languge of Classical India
Sindhi	North-west frontier of the Indian sub-continent

'GyanNidhi' which stands for 'Knowledge Resource' is parallel in 11 Indian languages , a project sponsored by TDIL, DIT, MC &IT, Govt of India

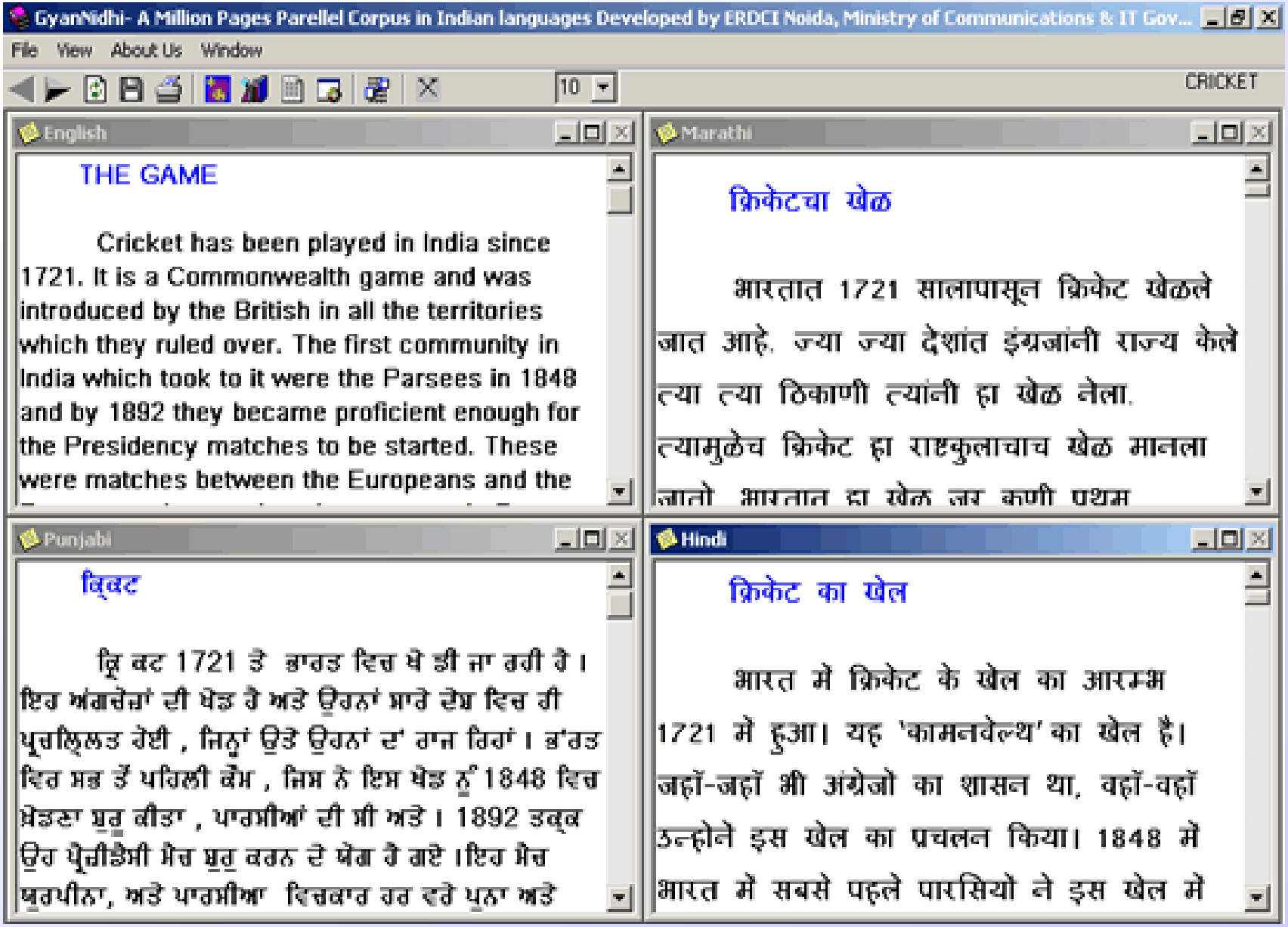
What GyanNidhi contains?

GyanNidhi corpus consists of text in English and 12 Indian languages (Hindi, Punjabi, Marathi, Bengali, Oriya, Gujarati, Telugu, Tamil, Kannada, Malayalam, Assamese, and Nepali).

It aims to digitise 1 million pages altogether containing at least 50,000 pages in each Indian language and English.

Prabandhika: Corpus Manager

- Categorization of corpus data in various user-defined domains
- Addition/Deletion/Modification of any Indian Language data files in HTML / RTF / TXT / XML format.
- Selection of languages for viewing parallel corpus with data aligned up to paragraph level
- Automatic selection and viewing of parallel paragraphs in multiple languages
  - Abstract and Metadata
  - Printing and saving parallel data in Unicode format


GyanNidhi- A Million Pages Parellel Corpus in Indian languages Developed by ERDCI Noida, Ministry of Communications & IT Gov...
File View About Us Window
10
CRICKET
English
THE GAME
Cricket has been played in India since 1721. It is a Commonwealth game and was introduced by the British in all the territories which they ruled over. The first community in India which took to it were the Parsees in 1848 and by 1892 they became proficient enough for the Presidency matches to be started. These were matches between the Europeans and the
Marathi
क्रिकेटचा खेळ
भारतात 1721 सालापासून क्रिकेट खेळले जात आहे. ज्या ज्या देशांत इंग्रजांनी राज्य केले त्या त्या ठिकाणी त्यांनी हा खेळ नेला. त्यामुळेच क्रिकेट हा राष्ट्रकुलाचाच खेळ मानला
Punjabi
ਕ੍ਰਿਕਟ
ਕ੍ਰਿਕਟ 1721 ਤੋਂ ਭਾਰਤ ਵਿਚ ਖੇਡੀ ਜਾ ਰਹੀ ਹੈ। ਇਹ ਅੰਗਰੇਜ਼ਾਂ ਦੀ ਖੇਡ ਹੈ ਅਤੇ ਉਹਨਾਂ ਸਾਰੇ ਦੇਸ਼ ਵਿਚ ਵੀ ਪ੍ਰਚਲਿਤ ਹੋਈ , ਜਿਨ੍ਹਾਂ ਉਤੇ ਉਹਨਾਂ ਦਾ ਰਾਜ ਕਿਹਾ। ਭਾਰਤ ਵਿਚ ਸਭ ਤੋਂ ਪਹਿਲੀ ਕੌਮ , ਜਿਸ ਨੇ ਇਸ ਖੇਡ ਨੂੰ 1848 ਵਿਚ ਖੇਡਣਾ ਸ਼ੁਰੂ ਕੀਤਾ , ਪਾਰਸੀਆਂ ਦੀ ਸੀ ਅਤੇ । 1892 ਤਕ ਉਹ ਪ੍ਰੈਜ਼ੀਡੈਂਸੀ ਮੈਚ ਸ਼ੁਰੂ ਕਰਨ ਦੇ ਯੋਗ ਹੋ ਗਏ ।ਇਹ ਮੈਚ ਯੂਰਪੀਨਾ, ਅਤੇ ਪਾਰਸੀਆ ਵਿਚਕਾਰ ਹਰ ਵਰੇ ਖੇਡਾ ਅਤੇ
Hindi
क्रिकेट का खेल
भारत में क्रिकेट के खेल का आरम्भ 1721 में हुआ। यह 'कामनवेल्थ' का खेल है। जहाँ-जहाँ भी अंग्रेजो का शासन था, वहाँ-वहाँ उन्होने इस खेल का प्रचलन किया। 1848 में भारत में सबसे पहले पारसियो ने इस खेल में

VISHLESHIKA

A software tool for conducting detailed Statistical Analysis of Text of Hindi language and adaptable to other Indian languages.

Statistics

- Sentence statistics
- Word statistics
- Cluster/conjunct statistics
- Character statistics
- Relative frequencies of Speech Sounds in Indian languages
- Extraction of phonetically rich sentences

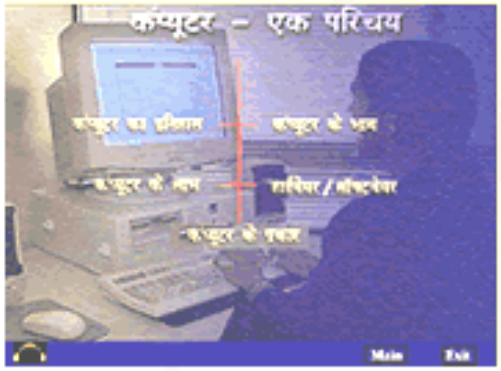
Character statistics

	Hindi		Punjabi		Marathi		Kannada	
Velars	क, क़, ख,ख़, ग,ग़, घ, ङ	17.8	ਕ, ਖ,ਖ਼, ਗ, ਗ਼, ਘ, ਙ	13.7	क, क़,ख,ख़, ग,ग़, घ, ङ	11.7	ಕ, ಖ, ಗ, ಘ, ಙ	13.6
Pre-palatal	च, छ, ज, ज़, झ, ञ,	6.1	ਚ, ਛ, ਜ, ਜ਼, ਝ, ਞ,	6.7	च, छ, ज, ज़, झ, ञ,	8.6	ಚ, ಛ, ಜ, ಝ, ಞ,	1.8
Retro flexes	ट, ठ, ड, ढ, ण	3.6	ਟ, ਠ, ਡ, ਢ, ਣ	9.0	ट, ठ, ड, ढ, ण	7.9	ಟ, ಠ, ಡ, ಢ, ಣ	8.2
Dentals	त, थ, द, ध, न,ऩ	18.2	ਤ, ਥ, ਦ, ਧ, ਨ,ਨ਼	23.1	त, थ, द, ध, न,ऩ	20.3	ತ, ಥ, ದ, ಧ, ನ	27.4
Labials	प, फ, फ़, ब, भ, म	14.4	ਪ, ਫ, ਫ਼, ਬ, ਭ, ਮ	11.6	प, फ, फ़, ब, भ, म	10.5	ಪ, ಫ, ಫ಼, ಬ, ಭ, ಮ	10.7
Semivowels	य, र, ऱ, ल, ळ, ऴ, व	21.8	ਯ, ਰ, ਲ, ੜ, ਵ	18.7	य, र, ऱ, ल, ळ, ऴ, व	28.3	ಯ, ರ,ಱ, ಲ, ಳ, ವ	29.7
Sibilants	श, ष, स	9.0	ਸ਼, ਸ,	7.4	श, ष, स	7.9	ಶ, ಷ, ಸ	6.5
Glottal	ह	8.1	ਹ	9.9	ह	4.9	ಹ	2.2
Flaps	ड़, ढ़	1.0						

The results from Table above show that in Kannada the occurrence of Dental consonants in much higher than in Hindi while in contrary usage of Glottal consonant in Hindi and Punjabi is much higher than in Marathi and Kannada.

हिंदी में कंप्यूटर प्रशिक्षण

- सीखने में आसान और बातचीत की शैली में
- अपनी क्षमता और समय से सीखें



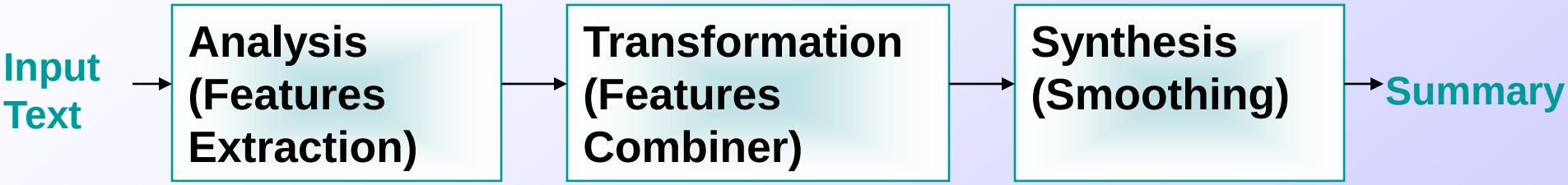
कंप्यूटर - एक परिचय

- कंप्यूटर का इतिहास
- कंप्यूटर के भाग
- कंप्यूटर के लाभ
- हार्डवेयर / सॉफ्टवेयर
- कंप्यूटर के प्रकार

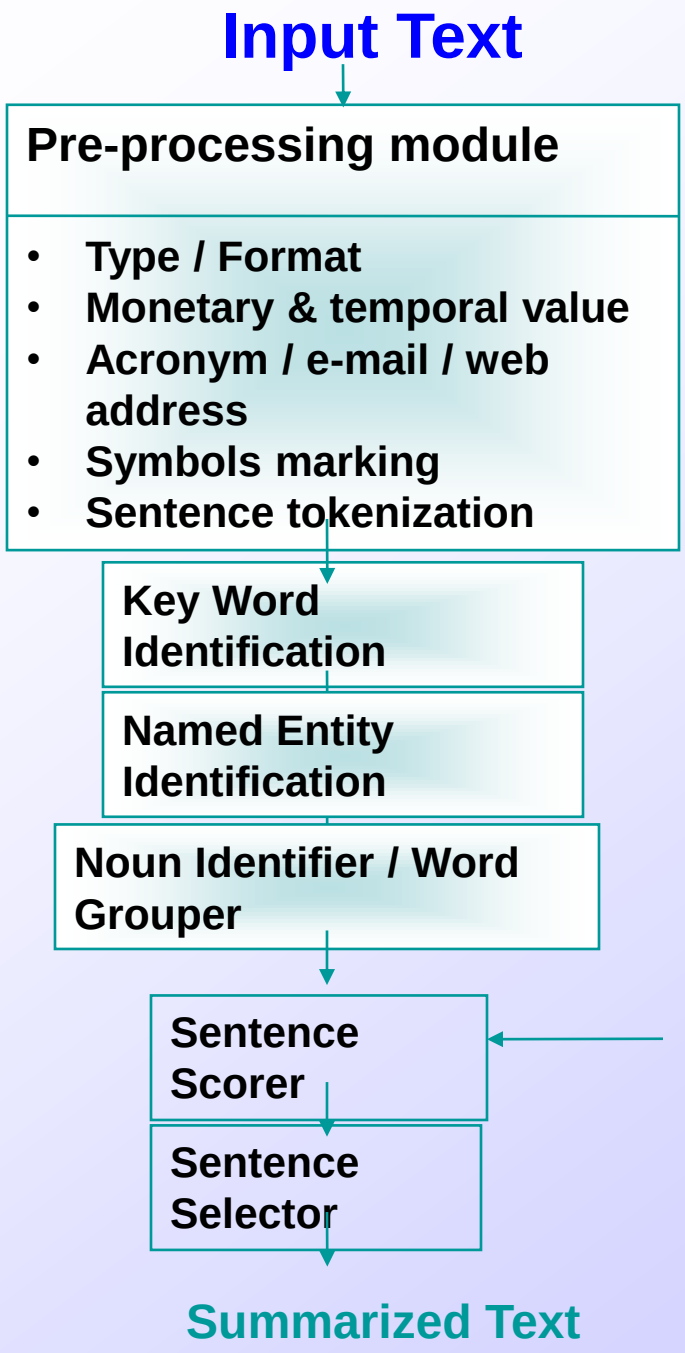
हिंदी सॉफ्टवेयर पैकेज पर कैसे कार्य करें

- वर्ड प्रोसेसिंग
- माइकोसॉफ्ट पावरपॉइन्ट
- ई मेल
- माइकोसॉफ्ट एक्सल
- इंटरनेट

Text Summarization : Broad Level Block Diagram



Text Summarization : Block Diagram



Heuristics Information

- Cue Phrases / Stigma
- Position & Format Information
- Title, Key word etc

Putting it Together: Linear Feature Combination

$$Weight(U) := \alpha * Location(U) + \beta * FixedPhrase(U) + \\ + \chi * ThematicTerm(U) + \delta * AddTerm(U)$$

U **is a text unit such as a sentence, *Greek letters* denote tuning parameters**

- **Location Weight assigned to a text unit based on whether it occurs in initial, medial, or final position in a paragraph or the entire document, or whether it occurs in prominent sections such as the document's introduction or conclusion**

- **FixedPhrase Weight assigned to a text unit in case fixed-phrase summary cues occur**

- **ThematicTerm Weight assigned to a text unit due to the presence of thematic terms (e.g., *tf.idf* terms) in that unit**

- **AddTerm Weight assigned to a text unit for terms in it that are also present in the title, headline, initial para, or the user's profile or query**

Tools/utilities/data for Summarization

- List of Stop words for Hindi
- Corpus of text in UNICODE (Scientific/News documents)
- Word Frequency count (Concordance tool after incorporating stemmer)
- Sentence Marker
- List of Cue phrases / Stigma Phrases in Hindi
- Stemming Algorithm implementation for Hindi to cover all inflections of single word for accurate frequency analysis and sentence scoring

Scoring of sentences is based on:

- Document Analysis (format, title, Heading, Paragraph, Position (Location))
- Presence of Key word/ stigma words/ Indicative phrases
- Identifying elaboration (redundancy) through marking text such as " such as , e.g., for example")

Mega Centre Digital Library

Objective

- Content Digitization (Scanning, Cleaning, Preservation and OCR)
- Tools Development

Future Plans of activities:
In- house Projects and International collaborative Efforts

- Collaborative project with A-Star …Contd.
- Development / improvement of technology systems for Hindi speech.
- Initiated collecting and transcribing conversational speech and broadcast news of Hindi and Indian English .
- Inter-institutional projects on Machine translation,multi-lingual resources ,OCR etc. in consortium mode.
- The institutions – Bengali (ISI/C-DAC Kolkata), Hindi(C-DAC Noida/TIFR), Indian English (C-DAC Pune, TIFR), Tamil (IITM/IISc, Bangalore), Telugu (IIIT /UoH Hyderabad) and Oriya (Utkal Univ, Bhubneshwar) have been proposed for developing suitable corpora and technologies.

Possible Collaboration with LDC

- Sharing of the linguistic resources and speech data collections already developed at CDAC, other institutions in India for use by Academic Institutions and Industries for Proto-type experiments.
- Joint Collaboration on New database development for Speaker and Language recognition development.
- Speaker/Language recognition system evaluation in collaboration with NIST.
- Setting up a joint transcription project between LDC and CDAC.
- Standards to be evolved for Evaluation of Systems

References

[1] Technology Development in Indian Languages Portal *www.tdil.mit.gov.in*

[2] S S Agrawal, K Samudravijaya, Karunesh Arora, “Text and Speech Corpora Development in Indian Languages”, *Proceedings of ICSLT-O-COCOSDA* 2004 New Delhi, India

[3]*Asia-Pacific Association for Machine Translation Journal, Special Issue*, MT Summit 2005, Phuket, Thailand.

[4] Ed. S S Agrawal et al, *Proc Intl. Symposium on Speech Technology and Processing Synthesis and O-COCOSDA-2004, vol II,* Tata McGraw Hill, Nov. 17-19,2004, New Delhi

[5] Ed. RMK Sinha et al, *Proc Intl. Symposium on Machine Translation NLP and TSS 2004, vol I* Tata McGraw Hill, Nov. 17-19,2004, New Delhi.

[6] Ed. K. Samudravijaya et al., Proc. Work on Spoken Language Processing, TIFR & ISCA, Jan 9-11, 2003, Mumbai.

THANKS