

Human Language Computing in Indian Languages - A Holistic Perspective



Swaran Lata

Country Manager , W3C India

Director & Head , TDIL Programme , Dept of Information Technology ,

Govt.of India

E-mail : slata@mit.gov.in

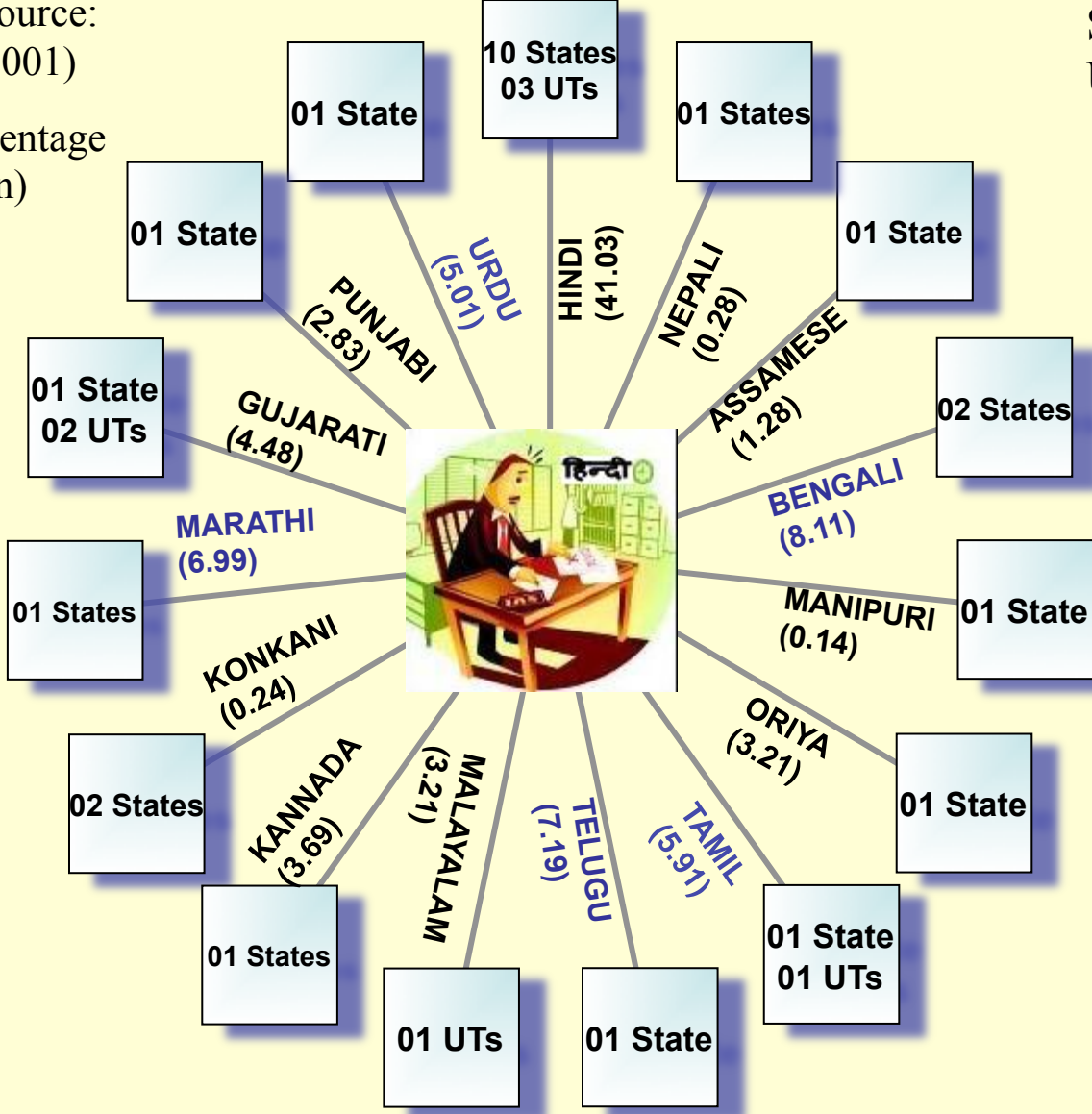
Organization of presentation:

- Languages of India and its distribution
- Technology Development for Indian Languages Programme
- Phases of TDIL Programme
- Paradigm Shift – Consortium mode projects
- Linguistic Resources developed
- Standardization Efforts
 - Core
 - Linguistic Resources
- Testing and Evaluation Initiatives
- Possible Collaborations with EU Programme
- Future Directions

Languages of India

- **Total Population:**
1,028,737,436 (Source: Census of India 2001)
- Language's (Percentage to total population)

INDIA
STATES: 28
UT: 07

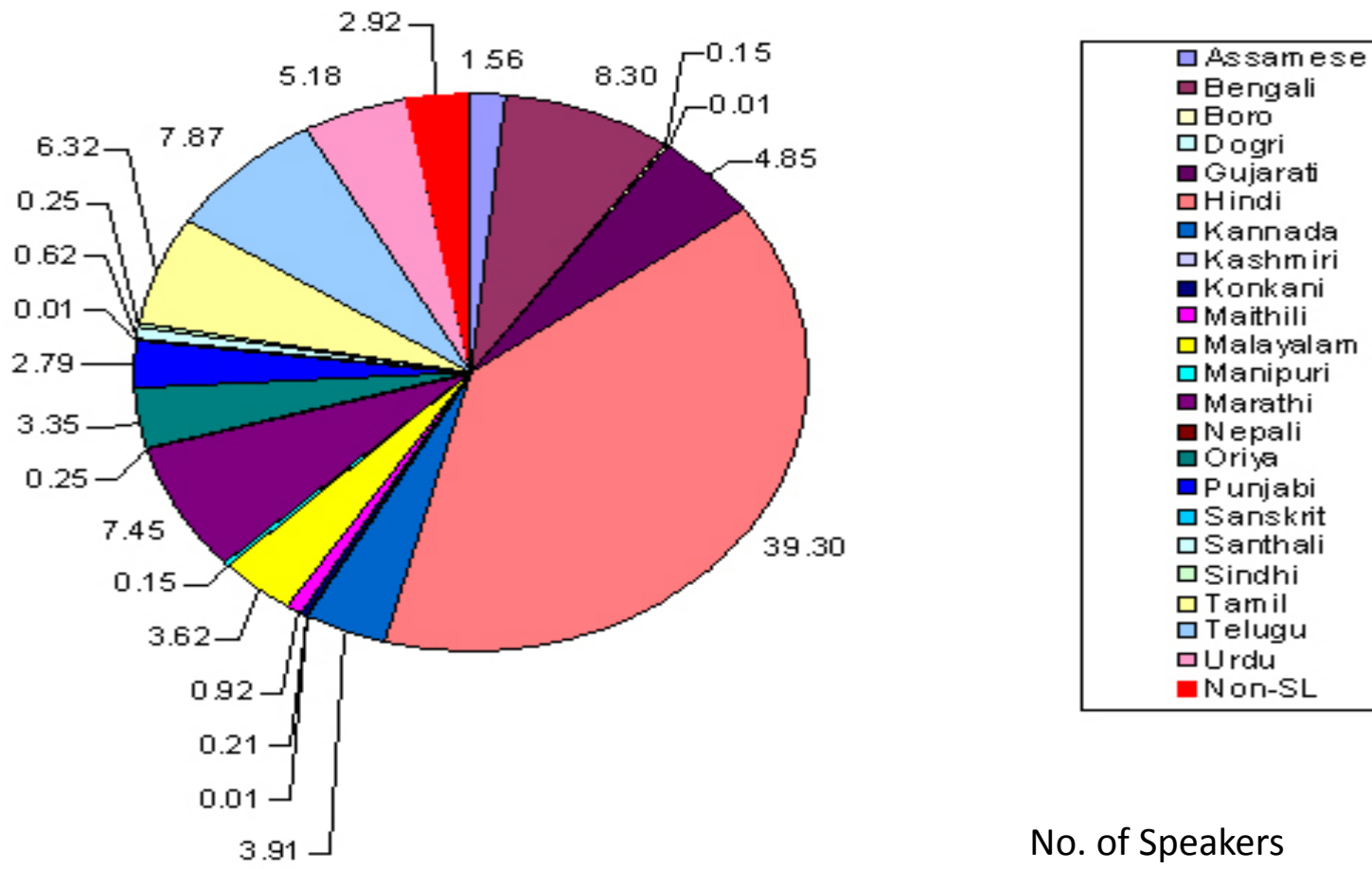


Linguistic Scenario in India

Source – Census 2001, India

Language	Speakers	Percentage to total population	State(s)
Assamese	13,168,484	1.28	Assam
Bengali	83,369,769	8.11	Andaman & Nicobar Islands, Assam, Tripura, West Bengal
Bodo	1,350,478	0.13	Assam
Dogri	2,282,589	0.22	Jammu and Kashmir
Gujarati	46,091,617	4.48	Dadra and Nagar Haveli, Daman and Diu, Gujarat
Hindi	422,048,642	41.03	Andaman and Nicobar Islands, Arunachal Pradesh, Bihar, Chandigarh, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand
Kannada	37,924,011	3.69	Karnataka.
Kashmiri	5,527,698	0.54	Jammu and Kashmir
Konkani	2,489,015	0.24	Goa, Karnataka, Maharashtra, Kerala
Maithili	12,179,122	1.18	Bihar
Malayalam	33,066,392	3.21	Kerala, Andaman and Nicobar Islands, Lakshadweep, Puducherry
Manipuri (also Meetei (Mayak)	1,466,705	0.14	Manipur
Marathi	71,936,894	6.99	Maharashtra, Goa, Dadra & Nagar Haveli, Daman and Diu, Madhya Pradesh, Karnataka
Nepali	2,871,749	0.28	Sikkim, West Bengal, Assam
Oriya	33,017,446	3.21	Orissa
Punjabi	29,102,477	2.83	Chandigarh, Delhi, Haryana, Punjab
Sanskrit	14,135	Negligible	Heritage Language
Santhali	6,469,600	0.63	Santhal tribals of the Chota Nagpur Plateau (comprising the states of Bihar, Chhattisgarh, Jharkhand, Orissa)
Sindhi	2,535,485	0.25	Non-regional language.
Tamil	60,793,814	5.91	Tamil Nadu, Andaman & Nicobar Islands, Puducherry;
Telugu	74,002,856	7.19	Andaman & Nicobar Islands, Andhra Pradesh, Puducherry;
Urdu	51,536,111	5.01	Jammu and Kashmir, Andhra Pradesh, Delhi, Bihar, Uttar Pradesh

Scheduled v/s Non-Scheduled Languages in India

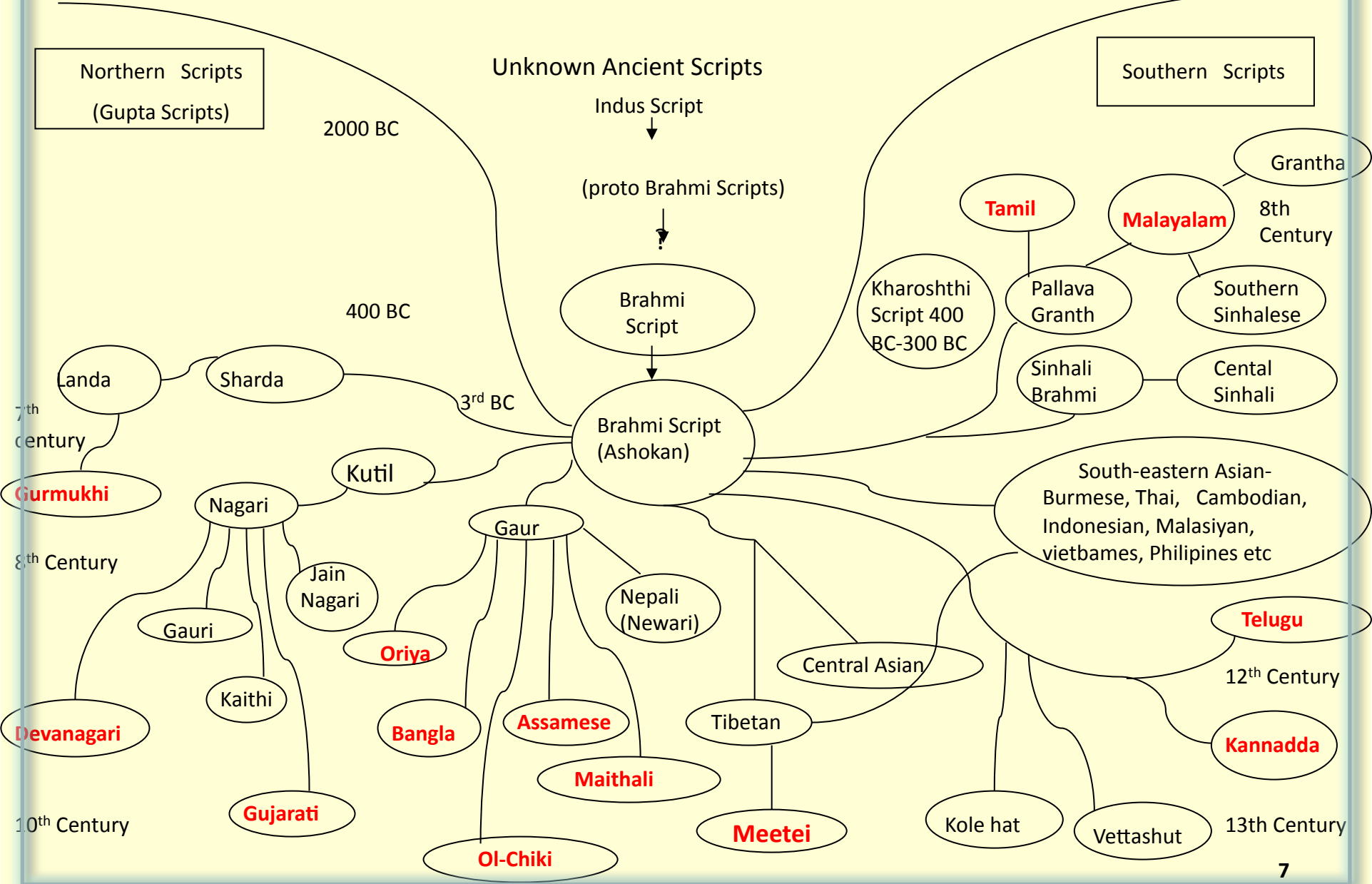


No. of Speakers

Official Indian Languages & Scripts

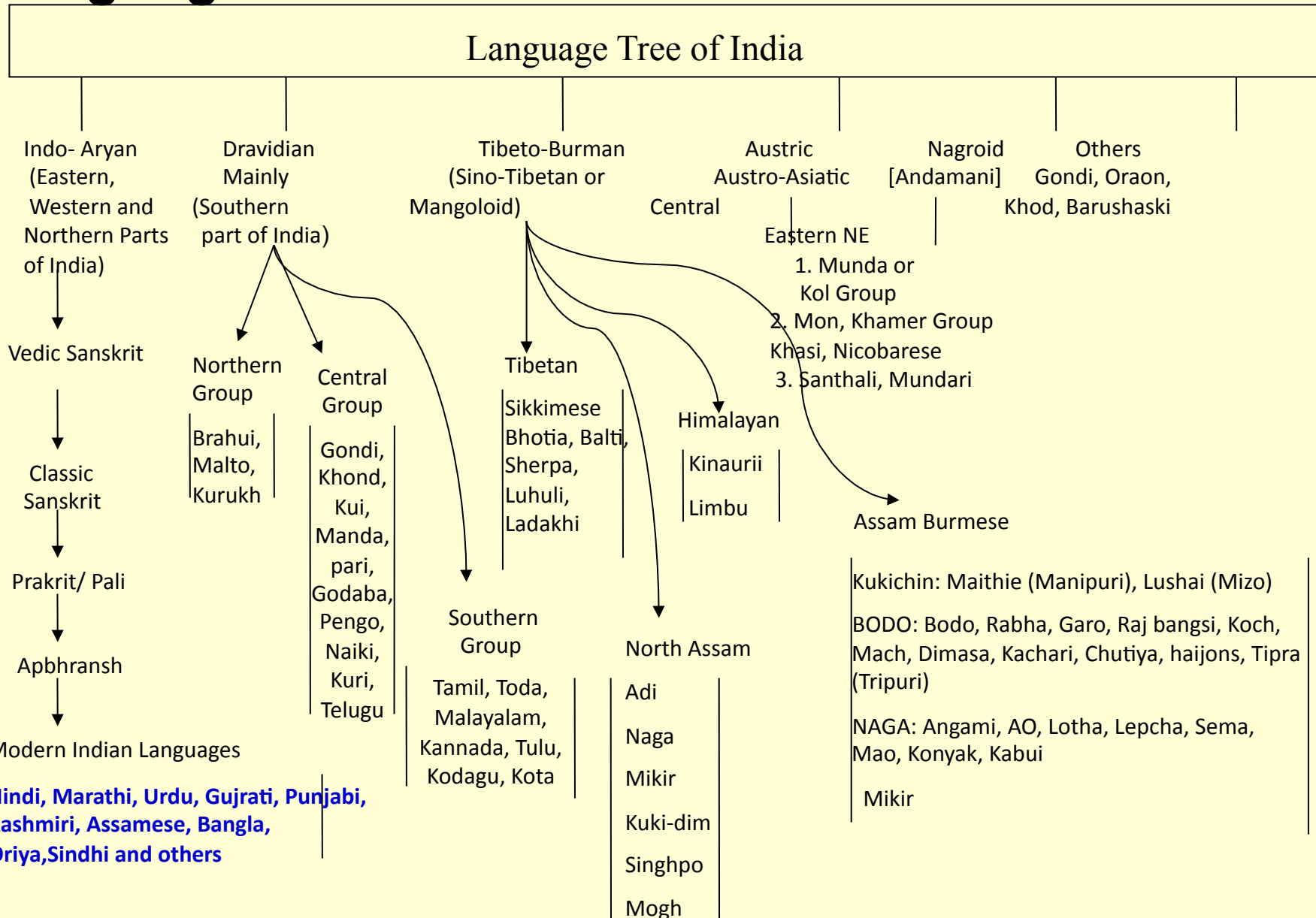
Sl. No.	Language	Script
1.	Hindi	Devanagari
2.	Sanskrit	Devanagari
3.	Marathi	Devanagari
4.	Konkani	Devanagari
5.	Nepali	Devanagari
6.	Maithili	Devanagari
7.	Sindhi	Devanagari
8.	Bodo	Devanagari
9.	Dogri	Devanagari
10.	Bengali	Bengali
11.	Assamese	Bengali
12.	Manipuri	Bengali, Meetei (Mayak)
13.	Gujarati	Gujarati
14.	Kannada	Kannada
15.	Malayalam	Malayalam
16.	Oriya	Oriya
17.	Punjabi	Gurmukhi
18.	Tamil	Tamil
19.	Telugu	Telugu
20.	Urdu	Arabic
21.	Santhali	Ol-Chiki, Devanagai,
22.	Kashmiri	Perso-Arabic, Devanagari

Major Scripts and Corresponding Languages in India



Languages of India

.....INDIA: A Primer



States

Languages

Chhattisgarh	Himachal Pradesh	Arunachal Pradesh	Karnataka	Kerala	Madhya Pradesh
Hindi	Hindi	Assamese	Kannada	Malayalam	Hindi
Maharashtra	Manipur	Mizoram	Nagaland	Orissa	Punjab
Marathi	Manipuri (Meitei)	Mizo, English	English	Oriya	Punjabi
Rajasthan	Sikkim	Tamil Nadu	Tripura	Andhra Pradesh	Chandigarh
Hindi	Nepali, English	Tamil	Bengali	Telugu	Punjabi
			English, Kokborok	Urdu	Hindi
Goa	Gujarat	Haryana	Jharkhand	Lakshadweep	Meghalaya
Konkani	Gujarati	Hindi	Hindi	Malayalam	English
Marathi	Hindi	Punjabi	Santhali	English	Khasi, Garo
Uttar Pradesh	West Bengal	Assam	Bihar	Dadra and Nagar Haveli	Daman and Diu
Hindi	Bengali	Assamese	Maithli	Gujarati	Gujarati
Urdu	Nepali	Bengali	Hindi	Marathi	English
		Bodo	Urdu	Hindi	Marathi
Delhi	Jammu and Kashmir	Puducherry	Uttarakhand	Andaman and Nicobar Islands	
Hindi	Urdu	Tamil	Hindi	Hindi	
Punjabi	Kashmiri	Malayala	Sanskrit	Bengali	
Urdu	Dogri	Telugu	Urdu	Tamil, Telugu	

TDIL Work Profile and Achievements

Introduction: TDIL Vision & Objectives

- ▶ Vision
Enabling masses to build knowledge society.

- ▶ Mission
Communicating without language barrier & moving up the knowledge chain.

- ▶ Objectives
 - To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access multilingual knowledge resources/content.

 - To consolidate technologies thus developed for Indian languages and integrate these towards wider proliferation and usage.

 - To promote collaborative development of futuristic technologies such leading to innovative products and services

Why Indian Language Technology is Important

- India being multilingual country needs software resources to be available in multiple language so that all linguistic communities take benefit out of it.
- It helps preserve Indian languages and culture
- It pushes employment and growth in India
- Rest of the world can customize their products for Indian market.
- It helps to increase e-development Index for transition into developed nation and an empowered society

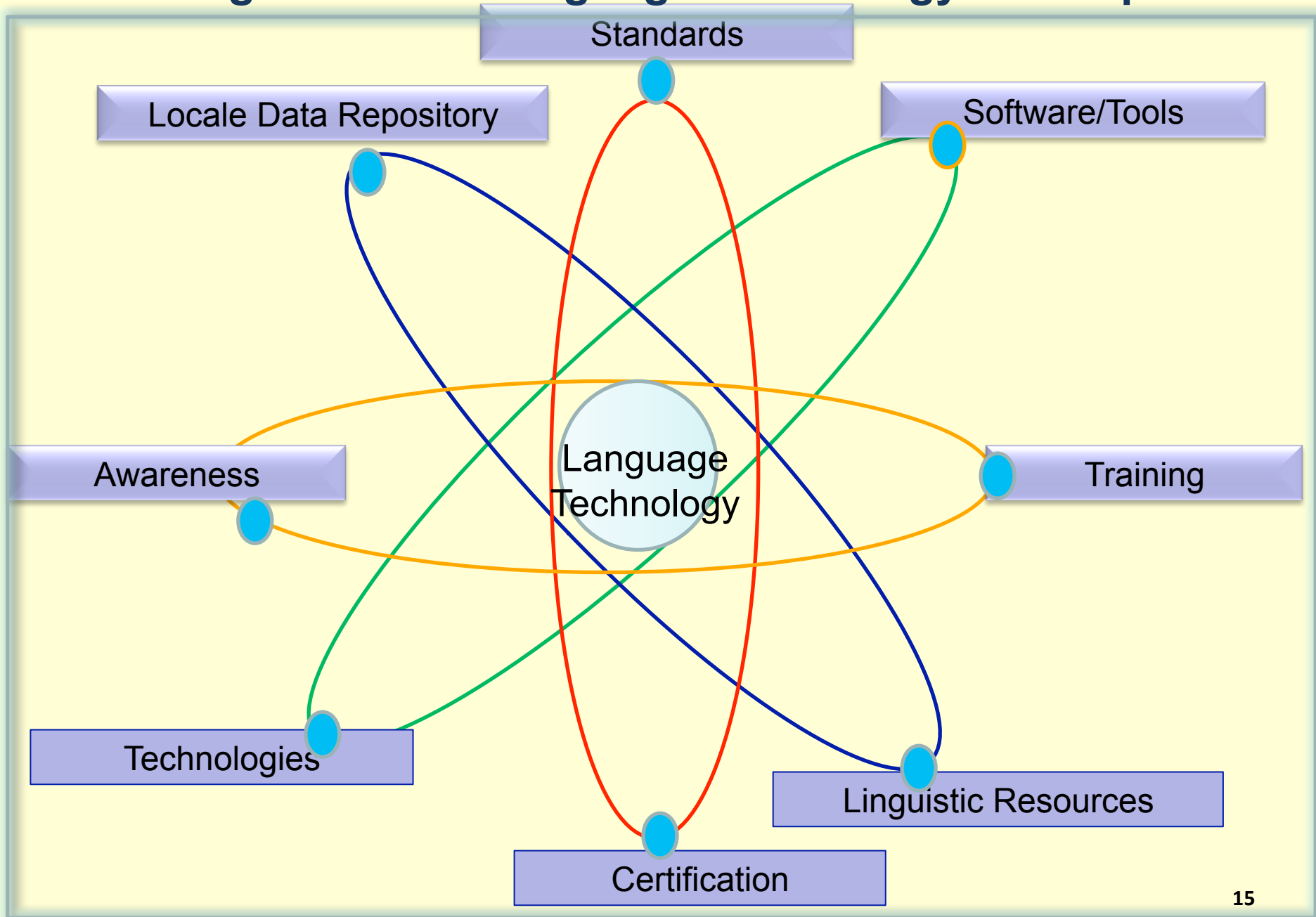
Local Language Interface – Not a desirable but An essential Component

- The success of increased mobile and broadband services hinges upon effective delivery of the citizen centric applications to rural masses.
- Since most of the citizens communicate in their local languages – Local Language Interface to G2C solutions at CSC is essential
- Hosting of content in local languages helps citizens to interact in a better way in today's knowledge society
- Thus , Indic Language Interface to services and applications is
 - “Not a desirable but An essential Component”

Thus the Role of Technology Development for Indian Languages Programme is very crucial and strategically important in:

- Developing and Bringing out key enabling technologies in 22 constitutionally recognized in Indian Languages
- Proliferation of Indian Language Technology to wider section of society
- Localization initiatives -
 - Promote localization industry
 - Support Indian languages on Indian ICT industry products/ solutions
 - Enable localization on Multinational product

Building Blocks of Language Technology Development



Phases of Language Technology development

- Seeding Phase : 1991-1995
 - TDIL programme established in the year 1991
 - Some linguistic resources such as corpora developed
 - NLP training programme for Computer Scientists and linguists
 - Some stand-alone language learning tools have also been developed
 - Exploratory Work in the area of NLP
- Exploratory Phase : 1995-2000
 - Development of Proof –of –concept Machine Translation System for English to Indian Languages and Indian Languages (Angla-Bharti) to Indian Languages (Anusaraka) systems have been developed.
 - Laboratory model of font dependent Optical Character Recognition in Hindi
 - Text-to-Speech for Hindi

◆ Catch-up Phase :2000-2004

The TDIL programme gathered momentum by establishing **13 Resource Centres** for Indian Languages Technology Solutions (RCILTS) and **10 CoIL-Net Centres**.

Resource Centres for Indian Languages Technology Solutions (RCILTS)

- The objective was to proliferate this activity to a large number of institutions across the country with the specific mandate for a language or a group of languages.
- Under this project, these centres have developed several important tools, linguistic resources and technologies for Indian language support
- Many of these tools are now being modified and upgraded to be released in public domain under National Roll-Out Project.

COIL-Net Centre:

- ◆ The objective was to develop Localized Content in Hindi Speaking states for enhancement of IT proliferation
- ◆ Initially there was minimal contents in Hindi with the initiation of Coil-net project Indian languages content have been generated
- ◆ E- content of approximately 16000 HTML & Dynamic pages in the domains of health, education, tourism and agri-business have been developed. Content on the eminent personalities, tourist places, classical work, and cultural heritage information on these regions have been developed.
- ◆ The developed content is uploaded on the internet at the website <http://tdil.mit.gov.in>.
- ◆ National Train Enquiry website localized in Hindi by CDAC. <http://www.trainenquiry.com> . It provides train tracking information.

▶ **Product Development and Proliferation Phase :2005-onwards**

- A '**Roadmap for Language Technology Development in India**' was evolved-to formulate short-term & long-term mission plan and strategy for development of Language Technologies in India.
- The Focus is to synergize development efforts and Develop deployable products
- **National Roll-Out Programme and Six Mission Mode Projects** have been initiated to facilitate Speedy Development & Availability of the Language Technologies.

INDIAN LANGUAGE TECHNOLOGY RESEARCH CENTRES IN INDIA



With TDIL sustained efforts Language Technology Approx. 80-100 Research Centers have been spread across the country.

Proliferation of Indian Language Technology Products : National Roll-Out Plan

Objectives of the initiative

To facilitate Speedy Development & Availability of the Language Technologies.

Broad contents of the CD

- Common user's Toolkit – Content Creation Tools, DTP, Office Automation, Code Converters
- Productivity Tools – Spellchecker, Domain based Dictionaries, Transliteration.
- Alpha version of technologies such as OCR, Text to Speech, MAT, etc

Distribution channel for the CD

- Registered users of www.ildc.in web site of TDIL, DIT – through postal department.
- IT magazines, publications, etc.
- Schools, Government departments, etc.

Software tools and fonts for 22 Indian languages namely Assamese, Bangla, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Manipuri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu languages have been released in public domain.

Freely downloadable from Indian Language Data centre – <http://www.ildc.gov.in>

Approx: 4.3 million downloads and 1.0 million shipments

CDs containing Indian Language Software Tools



Software tools and fonts CD contents

Common user –

- Unicode compliant Open Type fonts,
- True Type Fonts,
- Keyboard driver,
- Fonts and storage code converter,
- Localized version of Bharateeya OO (Office Suite),
- Fire fox browser,
- Email client,
- Typing Tutor,
- Spellchecker,
- Dictionaries

Power User –

- Text to Speech system,
- Transliteration Tool,
- Optical Character Recognition

Screen shots of Localized Bharatiyaa Open Office - autocorrect

The screenshot displays the OpenOffice.org 1.1.4 interface with the autocorrect dialog box open. The dialog box is titled "भाषा के लिए प्रतिस्थापना और आक्षेप" (Language-specific replacements and exceptions) and is set to "भंगीजी (USA)". The "प्रतिस्थापना" (Replacements) tab is active, showing a list of words and their corrections. The "केवल टेक्स्ट" (Text only) checkbox is checked. The "मिटाओ" (Remove) button is highlighted.

प्रतिस्थापना	से
म.प्र.	मध्य प्रदेश
उ.प्र.	उत्तर प्रदेश
कु.	कुमार
कुच	कुछ
खबरे	खबरे
डा.	डाक्टर
तरका	तड़का
दिल्लि	दिल्ली
नम्स्कार	नमस्कार
पड़ोसह	पड़ोसी
प्रो.	प्रोफेसर
म.प्र.	मध्य प्रदेश
महेस	महेश
रमेश	रमेश
रात्रेश	रात्रेश
राजिव	राजीव
श्रीमाम	श्रीमान
समचार	समाचार
स्वगत	स्वागत

Buttons at the bottom of the dialog: ठीक (OK), रद्द करो (Cancel), सुहायता (Help), रिसेट (Reset).

Windows taskbar at the bottom shows the Start button, OpenOffice.org 1.1.4 window, Gist OT Typing Tool, and system tray with the date 24 and time 1:09 PM.

Screen shots of Localized Bharatiyaa Open Office - spreadsheet

The screenshot displays the OpenOffice.org 1.1.4 spreadsheet interface. The window title is "नाम रहित1 - OpenOffice.org 1.1.4". The menu bar includes "फाइल", "संपादन", "दृश्य", "जोड़ो", "रचना", "औजार", "विंडो", and "सहायता". The toolbar contains various icons for file operations, editing, and formatting. The spreadsheet area shows a bar chart with the following data:

Row	Series 1	Series 2	Series 3
Row 1	9	3	4.5
Row 2	2.5	9	9.5
Row 3	3	1.5	3.5
Row 4	4.5	9	6

The chart has a legend with three series: "सम 1" (blue), "सम 2" (maroon), and "सम 3" (yellow). The y-axis ranges from 0 to 10. The x-axis labels are "Row 1", "Row 2", "Row 3", and "Row 4". Above the chart, there is a line of text in Hindi: "नहीं। वज्रानका न पहल दावा किया था कि चंद्रमा पर लगभग 40 साल पहल पाना को जास्तत्व था।". The status bar at the bottom shows "पृष्ठ 1 / 1", "डिफॉल्ट", "100%", "INSRT", "STD", "HYP", and a taskbar with the Windows start button, "नाम रहित1 - OpenOff...", "Gist OT Typing Tool", and system tray icons including the time "2:13 PM".

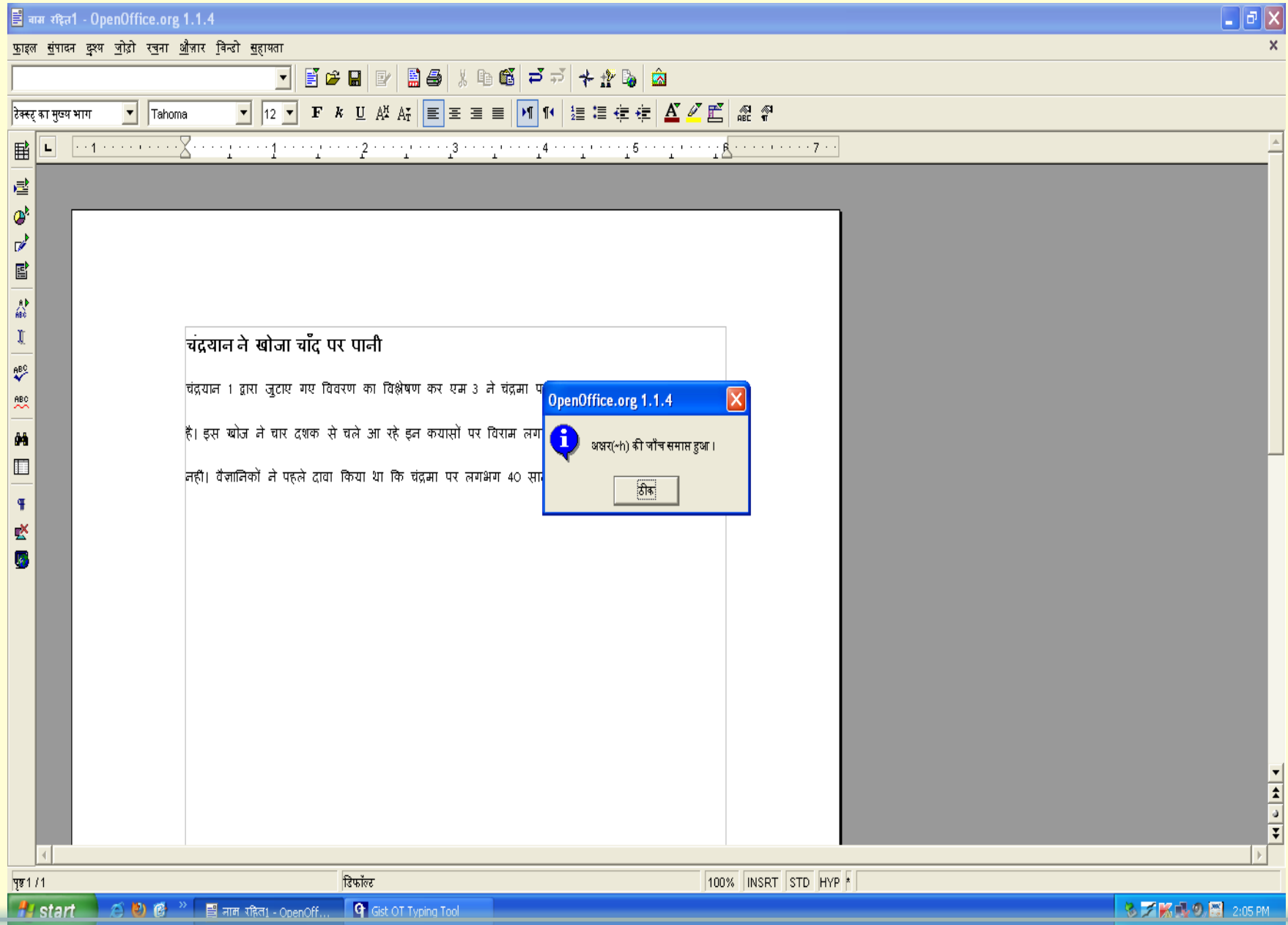
Screen shots of Localized Bharatiyaa Open Office - autosum

The screenshot shows the OpenOffice.org 1.1.4 interface with a spreadsheet. The spreadsheet has the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2	नाम	पिता का नाम	रोल न.	हिन्दी	गणित	योग(Total)									
3	विधि शर्मा	राकेश शर्मा	321540	86	89	175									
4	गीता रानी	मुकेश सिंह	321541	75	66	141									
5	रीति	ब्रजेश कुमार	321542	90	95	185									
6	गीतांजली यादव	विवेक यादव	321543	94	98	192									
7	खुशी पाल	वेद पाल	321544	77	68	145									
8	रेशमा	मुकेश सिंह	321545	62	48	=D8+E8									
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															

The spreadsheet interface includes a menu bar with options like 'फाइल', 'संपादन', 'दृश्य', 'जोड़ो', 'रचना', 'छाँटा', 'देरा', 'विन्डो', 'सहायता'. The toolbar contains various icons for file operations and editing. The status bar at the bottom shows 'शीट 1 / 3', 'डिफॉल्ट', '100%', 'INSRT', 'STD', '*', and 'जोड़=110'. The Windows taskbar at the very bottom shows the Start button and several open applications including 'Glist OT Typing Tool', 'sum_spreadsheet.PN...', and 'नाम रहित1 - OpenOff...'.

Screen shots of Localized Bharatiyaa Open Office (spell check tool)



Screen shots of Localized Bharatiyaa Open Office – pdf converter

The screenshot displays the OpenOffice.org 1.1.4 application window. The title bar reads 'नाम रहित1 - OpenOffice.org 1.1.4'. The menu bar includes 'फाइल', 'संपादन', 'दृश्य', 'जोड़ो', 'रचना', 'औजार', 'विन्डो', and 'सहायता'. The toolbar contains various icons for file operations and editing. The main window shows a document with the following text in Hindi:

चंद्रयान ने खोजा चाँद पर पानी
चंद्रयान 1 द्वारा जुटाए गए विवरण का
है। इस खोज ने चार दशक से चले आ
नहीं। वैज्ञानिकों ने पहले दावा किया थ

Below the text is a bar chart with a vertical axis from 0 to 10 and a horizontal axis labeled 'Row 1'. The chart has two bars: a purple bar at approximately 9.5 and a blue bar at approximately 3.5.

An 'Export' dialog box is open, showing the 'Save in' location as 'My Documents'. The file list includes folders like 'AdobeStockPhotos', 'Downloads', 'LG Electronics', 'My Music', 'My Pictures', 'OneNote Notebooks', 'SQL Server Management Studio', 'Updater', 'Visual Studio 2005', and 'Visual Studio 2008', along with a file named 'अ.pdf'. The 'File name' field contains 'लेख' and the 'File format' is set to 'PDF - Portable Document Format (.pdf)'. There are 'Save' and 'Cancel' buttons, and two checked options: 'सुरक्षित फाइल नाम विस्तार' and 'चुनाय'.

The status bar at the bottom shows 'पृष्ठ 1 / 1', 'डिफॉल्ट', '100%', 'INSRT', 'STD', 'HYP *', and the Windows taskbar with the 'start' button and several open applications including 'नाम रहित1 - OpenOff...', 'Gist OT Typing Tool', and 'untitled - Paint'. The system clock shows '2:17 PM'.

Screen shots of Localized Bharatiyaa Open Office – find & Replace

The screenshot shows the OpenOffice.org 1.1.4 interface in Hindi. The main window displays a presentation slide with the title "मुंबई ने बंगलौर को दो रन से हरा" (Mumbai defeated Bangalore by 2 runs). The slide content includes:

- चैम्पियंस लीग मुंबई इंडियंस ने बंगलौर को दो रन से हरा।
- इस जीत से मुंबई इंडियंस चैम्पियंस लीग टूर्नामेंट में बाहर होने से बच गईं।
- मुंबई इंडियंस के ब्रावो को **मैनन** आफ द मैच चुना गया. उन्होंने दो विकेट लिए और 29 रन भी बनाए।

The "Search & Replace" dialog box is open, showing the search term "मैनन" (Mannan) and the replacement term "मैन" (Man). The dialog box has the following fields and buttons:

- के लिए खोज: मैनन
- से प्रतिस्थापित करो: मैन
- विकल्प: केवल पूर्ण शब्द, पीछे की ओर, केस तुलना, केवल प्रचलित चुनाव, तुल्यता अन्वेषण
- Buttons: सब अन्वेषण करो, अन्वेषण, सब प्रतिस्थापित करो, प्रतिस्थापन, बन्द करो, सहायता

The status bar at the bottom shows the current slide is "स्लाइड 1" (Slide 1) and the zoom level is 100%.

Screen shots of Localized Bharatiyaa Open Office

The screenshot displays the OpenOffice.org 1.1.4 interface in Hindi. The main window shows a document with the following text:

चंद्रयान ने खोजा चाँद पर पानी
चंद्रयान 1 द्वारा जुटाए गए विवरण का
है। इस खोज ने चार दशक से चले आ
नहीं। वैज्ञानिकों ने पहले दावा किया था

Below the text is a bar chart with the following data:

Row	Value
Row 1	9
Row 2	3
Row 3	4.5

The 'क्षेत्र' (Field) dialog box is open, showing a list of fields:

- शर्तों के अन्तर्गत टेम्प्लेट
- Input list
- इनपुट क्षेत्र
- मैक्रो चलाओ
- स्थानधासक
- अक्षरों को मिलाओ
- छिपा हुआ टेम्प्लेट
- छिपा हुआ परिच्छेद

The dialog box also has tabs for 'लेखपत्र', 'रेफरेन्स', 'फंक्शन्स', 'DocInformation', 'वेरिफैबल', and 'लेखासंचय'. It includes fields for 'शर्त' (Condition) and 'Then'/'Else' clauses, and buttons for 'जोड़ो' (Add), 'बन्द करो' (Close), and 'सहायता' (Help).

The Windows taskbar at the bottom shows the Start button and several open applications: 'नाम रहित1 - OpenOff...', 'Gist OT Typing Tool', and 'untitled - Paint'. The system clock shows 2:15 PM.

Screen shots of Localized Bharatiyaa Open Office –insert link

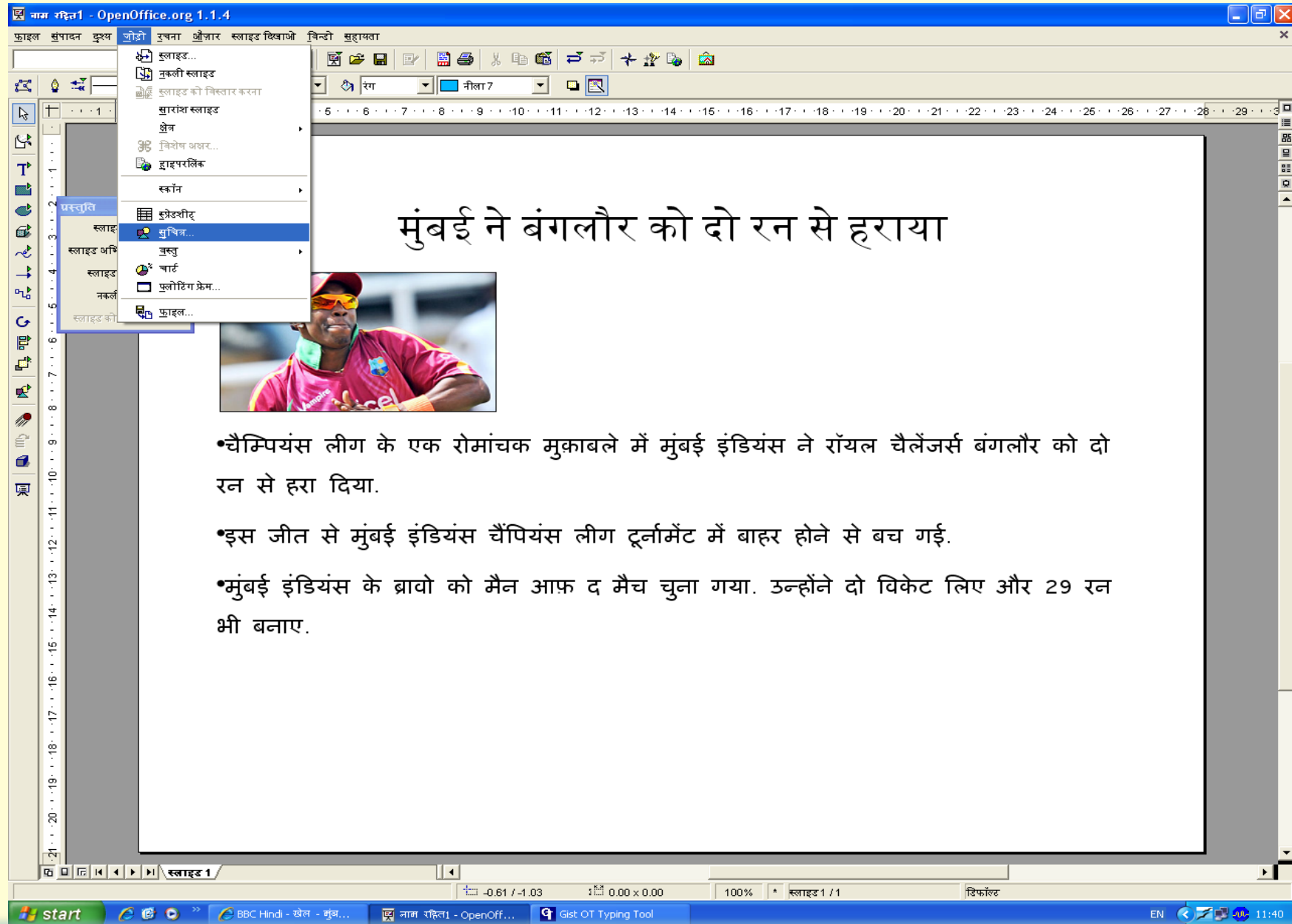
The screenshot displays the OpenOffice.org 1.1.4 interface in Hindi. The main window shows a presentation slide with the title "मुंबई ने बंगलौर को दो रन से हराया" (Mumbai defeated Bangalore by two runs). The slide content includes a photo of a cricketer and text: "चैम्पियंस रन से हरा", "इस जीत", "मुंबई इंडि रन भी बनाए. विस्तार से", "गलौर को दो", and "र और 29".

Overlaid on the slide is a "इंटरनेट लिंक" (Internet Link) dialog box. It contains the following fields and options:

- हाइपरलिंक वर्ग (Hyperlink Class): इन्टरनेट (Internet), ETP, Inetnet
- लुप्त (URL): bbc.co.uk/hindi/sport/2010/09/100920_mumbai_banglore_ac.shtml
- अधिक निर्धारण (Additional Settings):
 - फ्रेम (Frame):
 - रेकॉर्ड (Record): (विस्तार से)
 - नाम (Name):
- Buttons: लागू करो (Apply), बन्द करो (Cancel), सहायता (Help), पीछे (Back)


At the bottom of the screen, the Windows taskbar shows the Start button, system tray with the time 11:48, and several open applications including "BBC Hindi - खेल - मुंब...", "नाम रहित1 - OpenOff...", "Gist OT Typing Tool", and "find_presentation.PN...". The page number "31" is visible in the bottom right corner.

Screen shots of Localized Bharatiyaa Open Office



The screenshot shows the OpenOffice.org 1.1.4 interface in Hindi. The main window displays a slide with the following content:

मुंबई ने बंगलौर को दो रन से हराया



- चैम्पियंस लीग के एक रोमांचक मुकाबले में मुंबई इंडियंस ने रॉयल चैलेंजर्स बंगलौर को दो रन से हरा दिया.
- इस जीत से मुंबई इंडियंस चैम्पियंस लीग टूर्नामेंट में बाहर होने से बच गईं.
- मुंबई इंडियंस के ब्रावो को मैन आफ द मैच चुना गया. उन्होंने दो विकेट लिए और 29 रन भी बनाए.

The interface includes a menu bar with options like 'फाइल', 'संपादन', 'दृश्य', 'जोड़ी', 'रचना', 'औजार', 'स्लाइड दिखाओ', 'विन्डो', and 'सहायता'. A toolbar with various icons is visible below the menu bar. The status bar at the bottom shows the current slide as 'स्लाइड 1 / 1' and the zoom level as '100%'.

Consortium Approach- Paradigm shift in Language Technology Development:

- To bring out deployable products from core technology
- To address complex Indian Language technology issues
- To converge the expertise of the scientists / researchers as no single group may be in a position to develop the product - Putting Institutions Together
- Separating out the core engine from the language verticals and responsibility for core engine development at the Consortium leading institutions
- Language Verticals to be handled at different institutions in respective states
- Uniformity in approach as inherent modules of the system need to be integrated
- Involvement of Industry Partner for System Integration and Software Engineering perspective
- Once core engine is developed Industry partner may be involved to incorporate and fine tune the basic technology
- Standard Software Engineering Practices need to be invoked for product development -Industry Partner may join hand as consultant

Technologies Developed under consortium mode projects

- English to Indian Languages Machine Translation System
[6 Language Pairs: English to Hindi, Marathi, Bengali, Oriya, Tamil, Urdu.]
- Indian Languages to Indian Languages Machine Translation System
[9 Language Pairs: Telugu-Hindi, Hindi-Tamil, Urdu-Hindi, Kannada-Hindi, Punjabi-Hindi, Marathi-Hindi, Bengali-Hindi, Tamil-Telugu, Malayalam-Tamil]
- Cross-Lingual Information Access
[6 Languages : Hindi , Bengali, Tamil , Marathi , Telugu and Punjabi]
- Optical Character Recognition Systems
[10 Scripts: Bangla, Devnagari, Malayalam, Gujrati, Tamil, Telugu, Kannada, Oriya, Gurumukhi, Tibetan]
- On-line Handwriting recognition system [6 Scripts: Hindi , Bengali , Tamil , Telugu , Kannada and Malayalam]

Sample Outputs For English - Urdu

The screenshot displays the 'Indian Language Machine Translation System' interface within a Windows Internet Explorer browser. The address bar shows the URL: <http://tdil-dc.in/sampark/web/index.php/content/demotranslation>. The page features a header banner for the 'Indian Language Technology Proliferation and Deployment Centre (ILTP-DC)' with the motto 'सत्यमेव जयते' and the text 'भारतीय भाषा प्रौद्योगिकी प्रसरण एवं विस्तारण केन्द्र'. Below the banner, the page title is 'Sampark : Machine Translation among Indian Languages (Experimental Version)', funded by the TDIL Program, Department of Information Technology, Govt. of India, and developed by a Consortium of Institutions*. The interface includes a navigation menu with options like 'Text', 'Web Page', 'Document', 'Demo', 'Feedback', 'History', 'Help', and 'Duar Sampark Text'. A main instruction reads 'Select language pair and click translate button.'. A text input field contains the Urdu text: 'اجے پال چوبان نے ساتویں صدی سے اجمیر کی بنیاد رکھی -'. Below the input field, a dropdown menu is open, showing language pairs: 'Urdu to Hindi', 'Punjabi to Hindi', 'Hindi to Punjabi', 'Urdu to Hindi' (highlighted), 'Telugu to Tamil', 'Hindi to Telugu', and 'Tamil to Hindi'. A yellow 'Translate' button is positioned to the right of the dropdown. The browser's taskbar at the bottom shows the system tray with the date and time '16:24' and the language set to 'EN'. The Windows taskbar includes several open applications: 'Welcome in Indian ...', 'اردو یوب سرورگ - Win...', and 'ILMT1_out_pun.jpg ...'.

Sample Outputs For English - Bangla

Translated Outputs - Windows Internet Explorer
http://tdil-dc.in/eilmt/uploadText.do

File Edit View Favorites Tools Help

Translated Outputs

टी डी आई एल TDIL
भारतीय भाषा प्रौद्योगिकी प्रसारण एवं विस्तारण केन्द्र

Indian Language Technology Proliferation and Deployment Centre
ILTP-DC

भारतीय भाषा प्रौद्योगिकी प्रसारण एवं विस्तारण केन्द्र सत्यमेव जयते

English To Indian Languages Machine Translation System

You are : **cdac** Language: **Bengali** Home | Help | TdilHome

* This page can display maximum of 5 outputs at a time.

Input	Output	More..
Jaipur is the pink city of India	जयपुर भारत के गोलापी शहर है	

Final Output

Developed by Consortium of Institutions - C-DAC Pune, IIIT-Hyderabad, IIT-Bombay, IISc-Bangalore, Jadavpur University, C-DAC Mumbai, Utkal University, Banasthali Vidyapeeth, Amrita University, IIIT-Allahabad Experimental System.

© 2009-10 Department of Information Technology, MCIT, Govt of India

W3C XHTML 1.0 W3C CSS india.gov.in


Done Internet | Protected Mode: On 120% 15:36

Sample Outputs For English - Tamil

Translated Outputs - Windows Internet Explorer
http://tdil-dc.in/eilmt/uploadText.do

File Edit View Favorites Tools Help

Translated Outputs




Indian Language Technology Proliferation and Deployment Centre
ILTP-DC
भारतीय भाषा प्रौद्योगिकी प्रसरण एवं विस्तारण केन्द्र सत्यमेव जयते

English To Indian Languages Machine Translation System

You are : **cdac** Language: **Tamil** Home | Help | TdilHome

* This page can display maximum of 5 outputs at a time.

Input	Output	More..
Jaipur is the pink city of India	ஜெய்ப்பூர் இந்தியாவின் இளஞ்சிவப்புநிறம் நகரமாக இருக்கிறது	

Final Output

Developed by Consortium of Institutions - C-DAC Pune, IIIT-Hyderabad, IIT-Bombay, IISc-Bangalore, Jadavpur University, C-DAC Mumbai, Utkal University, Banasthali Vidyapeeth, Amrita University, IIIT-Allahabad Experimental System.

© 2009-10 Department of Information Technology, MCIT, Govt of India

W3C XHTML 1.0 W3C CSS india.gov.in

Internet | Protected Mode: On 120% 15:39

Indian Language to Indian Languages Machine Translation System

The screenshot shows a web browser window displaying the Indian Language Machine Translation System. The browser's address bar shows the URL: <http://tdil-dc.in/sampark/web/index.php/content/demotranslation>. The page features a header banner for the Indian Language Technology Proliferation and Deployment Centre (ILTP-DC) with the text: "Indian Language Technology Proliferation and Deployment Centre" and "भारतीय भाषा प्रौद्योगिकी प्रसरण एवं विस्तारण केन्द्र सत्यमेव जयते". Below the banner, the page title is "Sampark : Machine Translation among Indian Languages (Experimental Version)", funded by the TDIL Program, Department of Information Technology, Govt. of India, and developed by a Consortium of Institutions*. The interface includes a navigation menu with tabs: Text, Web Page, Document, Demo, Feedback, History, Help, and Duur Sampark Text. A message reads: "Select language pair and click translate button." and "Translation Completed." The main content area shows a Telugu input: "ఈ విధానికి నాలుగు ముఖాలు ఇంకా నాలుగు చేతులు ఉన్నాయి." and its Tamil translation: "இந்த விக்கிரகத்துக்கு நான்கு முகங்கள் இன்னும் நான்கு கைகள் இருக்கிறன .". Below the text area, there is a dropdown menu set to "Telugu to Tamil" and a "Translate" button. A list of instructions is provided: "Click Sampark Text tab for: Hindi to Punjabi, Punjabi to Hindi, Telugu to Tamil, Urdu to Hindi Translations" and "Click Duur Sampark Text tab for: Hindi to Telugu and Tamil to Hindi Translations". The Windows taskbar at the bottom shows the system tray with the time 16:30 and the date 12/16/2010.

Sample Outputs For Hindi -Punjabi

The screenshot displays the Sampark web application interface. At the top, there is a banner for the Indian Language Technology Proliferation and Deployment Centre (ILTP-DC) with the motto 'सत्यमेव जयते'. Below the banner, the application title 'Sampark : Machine Translation among Indian Languages (Experimental Version)' is shown, along with funding and development information. The interface includes a navigation menu with tabs for 'Text', 'Web Page', 'Document', 'Demo', 'Feedback', 'History', 'Help', and 'Duur Sampark Text'. A 'Type Text' input field contains the Hindi text 'नई दिल्ली भारत की राजधानी है |'. A 'Select keyboard' dropdown menu is set to 'd'. The output area shows the translated Punjabi text 'ਨਈ ਦਿੱਲੀ ਭਾਰਤ ਦੀ ਰਾਜਧਾਨੀ ਹੈ .'. A 'Translate' button is highlighted in yellow. At the bottom, there are instructions for using the application and a 'Notes' link.

Indian Language Technology Proliferation and Deployment Centre
ILTP-DC
भारतीय भाषा प्रौद्योगिकी प्रसरण एवं विस्तारण केन्द्र सत्यमेव जयते

Sampark : Machine Translation among Indian Languages (Experimental Version)
Funded by : TDIL Program, Department of Information Technology, Govt. of India
Developed by Consortium of Institutions*

Text Web Page Document Demo Feedback History Help Duur Sampark Text Welcome Cdac! TDIL-DC Home

Type Text Select keyboard d Translation Completed.

नई दिल्ली भारत की राजधानी है | नਈ ਦਿੱਲੀ ਭਾਰਤ ਦੀ ਰਾਜਧਾਨੀ ਹੈ .

Hindi to Punjabi Translate Suggest better translation View Tree

- Click Sampark Text tab for: Hindi to Punjabi, Punjabi to Hindi, Telugu to Tamil, Urdu to Hindi Translations
- Click Duur Sampark Text tab for: Hindi to Telugu and Tamil to Hindi Translations

Notes

Urdu-Hindi ILMT Translation through URL :- Urdu input

The screenshot displays the 'Sampark' web application for machine translation. The header features the TDIL logo and the text 'Indian Language Technology Proliferation and Deployment Centre' along with 'ILTP-DC' and the motto 'सत्यमेव जयते'. Below the header, the page title is 'Sampark : Machine Translation among Indian Languages (Experimental Version)'. The interface includes a navigation menu with options like 'Text', 'Web Page', 'Document', 'Demo', 'Feedback', 'History', 'Help', and 'Duur Sampark Text'. A status bar indicates 'Welcome Cdac! TDIL-DC Home' and 'Translation Completed.'. The main content area shows the Urdu input: 'اجے پال چوہان نے ساتویں صدی میں اجمیر کی بنیاد رکھی۔' and the corresponding Hindi output: 'अजय पाल चौहान ने सातवीं शताब्दी में अजमेर की नींव रखी।'. A dropdown menu is set to 'Urdu to Hindi' with a 'Translate' button. A footer note provides instructions: 'Click Sampark Text tab for: Hindi to Punjabi, Punjabi to Hindi, Telugu to Tamil, Urdu to Hindi Translations' and 'Click Duur Sampark Text tab for: Hindi to Telugu and Tamil to Hindi Translations'.

Welcome in Indian Language Machine Translation System - Windows Internet Explorer
http://tdil-dc.in/sampark/web/index.php/content/demotranslation

File Edit View Favorites Tools Help

Translated Outputs Welcome in Indian Lan...

टी डी आई एल TDIL
अज्ञान(अज्ञान)द्वारा-अज्ञान(अज्ञान)

Indian Language Technology Proliferation and Deployment Centre
ILTP-DC
भारतीय भाषा प्रौद्योगिकी प्रसरण एवं विस्तारण केन्द्र सत्यमेव जयते

सम्पर्क
Sampark : Machine Translation among Indian Languages (Experimental Version)
Funded by : TDIL Program, Department of Information Technology, Govt. of India
Developed by Consortium of Institutions*

Text Web Page Document Demo Feedback History Help Duur Sampark Text Welcome Cdac! TDIL-DC Home

Select language pair and click translate button. Translation Completed.

اجے پال چوہان نے ساتویں صدی میں اجمیر کی بنیاد رکھی۔
अजय पाल चौहान ने सातवीं शताब्दी में अजमेर की नींव रखी।

Urdu to Hindi Translate

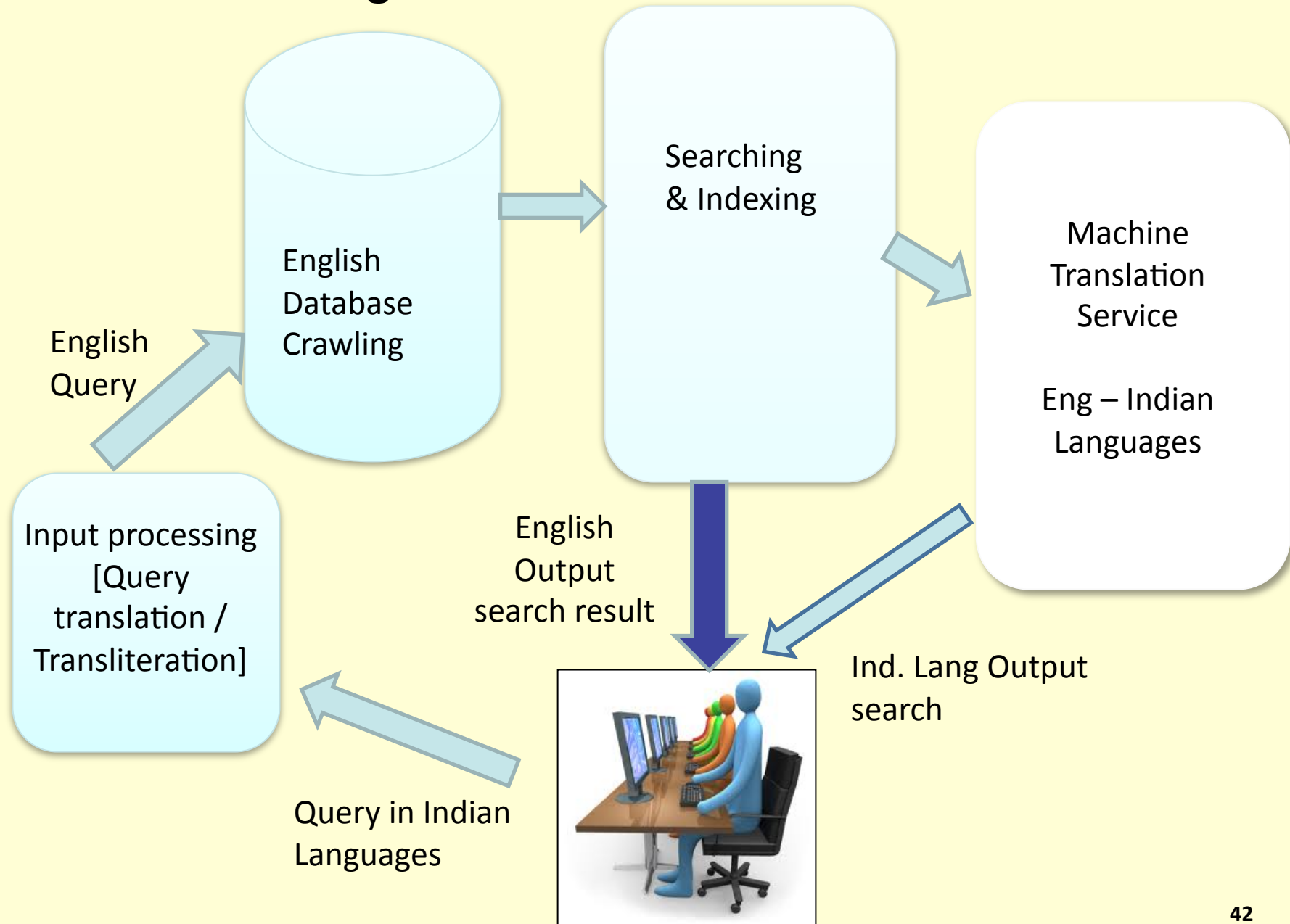
- Click Sampark Text tab for: Hindi to Punjabi, Punjabi to Hindi, Telugu to Tamil, Urdu to Hindi Translations
- Click Duur Sampark Text tab for: Hindi to Telugu and Tamil to Hindi Translations

Notes

Internet | Protected Mode: On 120%

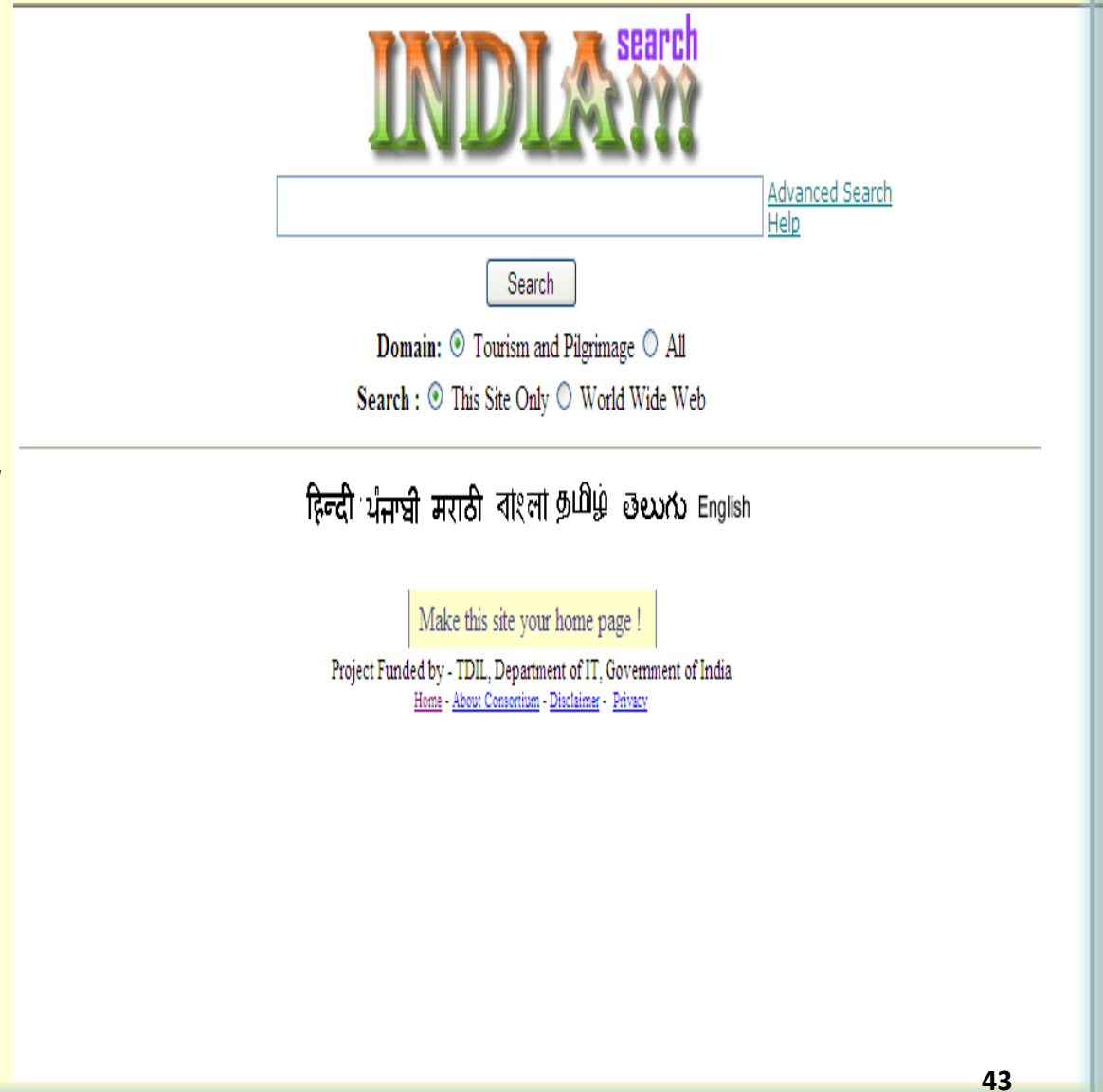
Welcome in Indian ... ILMT_urdu.jpg - Paint EN 16:26

CLIA integrated with Machine Translation



Cross-Lingual Information Access (CLIA)

- In CLIA, the input query is in one language and information is retrieved in another.
- The query language is one of Bangla, Hindi, Marathi, Punjabi, Tamil and Telugu.
- The retrieved documents are in English, Hindi or the language of the query.



The screenshot shows the homepage of the INDIA search engine. At the top, the word "INDIA" is written in large, colorful, 3D-style letters, with "search" in purple text to its right. Below this is a search input field with a "Search" button. To the right of the input field are links for "Advanced Search" and "Help". Below the search field, there are radio buttons for "Domain" (selected: "Tourism and Pilgrimage", unselected: "All") and "Search" (selected: "This Site Only", unselected: "World Wide Web"). A horizontal line separates the search area from the language selection area. Below the line, the text "हिन्दी पंजाबी मराठी बांग्ला தமிழ் తెలుగు English" is displayed. Below this is a yellow box with the text "Make this site your home page!". At the bottom, it says "Project Funded by - TDIL, Department of IT, Government of India" followed by links for "Home", "About Consortium", "Disclaimer", and "Privacy".

খোঁজ

খোঁজ : এই ইঞ্জিন দ্বারা অন্যান্য ইঞ্জিন দ্বারা

হিন্দি ইংরাজী বাংলা

পরিণাম 1 থেকে 10 মোট 1221 খোঁজ -

কলকাতা

যদিও ভারতের অন্যান্য মহানগরগুলির মতো নগরায়ণ সমস্যার অঙ্গ হিসাবে দারিদ্র, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে একেবারে [http://bn.efactory.pl/কলকাতা \(ভাঙার\) \(সারণ্য\)](http://bn.efactory.pl/কলকাতা_(ভাঙার)_সারণ্য)

কলকাতা মেট্রো

কলকাতার বিভিন্ন যানবাহন সম্পর্কে বহু অভিযোগ থাকলেও মেট্রো সম্পর্কে অভিযোগ খুবই কম। **কলকাতার মেট্রো** সাধারণত সঠিক [http://bn.efactory.pl/কলকাতার_মেট্রো \(ভাঙার\) \(সারণ্য\)](http://bn.efactory.pl/কলকাতার_মেট্রো_(ভাঙার)_সারণ্য)

কলকাতা মেট্রো

কলকাতার বিভিন্ন যানবাহন সম্পর্কে বহু অভিযোগ থাকলেও মেট্রো সম্পর্কে অভিযোগ খুবই কম। **কলকাতার মেট্রো** সাধারণত সঠিক [http://bn.efactory.pl/কলকাতা_মেট্রো \(ভাঙার\) \(সারণ্য\)](http://bn.efactory.pl/কলকাতা_মেট্রো_(ভাঙার)_সারণ্য)

কলকাতার ইতিহাস

Searching for: **কলকাতার ইতিহাস** ? **কলকাতা** সর্ববৃহৎ শহর ও প্রধান বন্দর হিসেবে বিরাজমান এ মহানগরীটি ১৯১১ মালের ডিসেম্বর [http://bn.efactory.pl/কলকাতার_ইতিহাস \(ভাঙার\) \(সারণ্য\)](http://bn.efactory.pl/কলকাতার_ইতিহাস_(ভাঙার)_সারণ্য)

কলকাতা - উইকিপিডিয়া

যদিও ভারতের অন্যান্য মহানগরগুলির মতো নগরায়ণ সমস্যার অঙ্গ হিসাবে দারিদ্র, পরিবেশ দূষণ ও যানজটের সমস্যা থেকে একেবারে [http://bn.wikipedia.org/wiki/কলকাতা \(ভাঙার\) \(সারণ্য\)](http://bn.wikipedia.org/wiki/কলকাতা_(ভাঙার)_সারণ্য)

উইকিপিডিয়া একটি মুক্ত বিশ্বকোষ

প্রধান পাতা
সম্প্রদায়ের প্রবেশদ্বার
সমসাময়িক ঘটনা
সাম্প্রতিক পরিবর্তনসমূহ
অন্যান্য যেকোনো পৃষ্ঠা
সহায়তা
দান করুন

▼ মূলপ্রসঙ্গটি
বই তৈরি করুন
PDF ডাউনলোড
স্থানীয় খোঁজ সংস্করণ

► সরাসরি

▼ অন্যান্য ভাষাসমূহ
Afrikaans
Aragonés
/Englisc
العربية
عصرى
অসমীয়া

কলকাতা

"কলিকাতা" এখানে পুননির্দেশ করা হয়েছে। অন্য ব্যবহারের জন্য, দেখুন কলিকাতা (স্বাক্ষর নিরসন)

কলকাতা (পূর্বনাম: **কলিকাতা**, ইংরেজি: Kolkata, পূর্বে **Calcutta** ^{কলকাতা}) ভারতের পূর্বাঞ্চলীয় রাজ্য পশ্চিমবঙ্গের রাজধানী, প্রধান বাণিজ্যকেন্দ্র এবং বৃহত্তম শহর। হুগলী নদীর পূর্ব তীরে অবস্থিত^[১] এই শহরের পৌরপ্রশাসকের জনসংখ্যা ৫০ লাখের কিছু বেশি। তবে কলকাতা ও তার পার্শ্ববর্তী জেলাগুলির অংশেবিশেষ নিয়ে গঠিত বৃহত্তর কলকাতার জনসংখ্যা ১ কোটি ৪০ লাখের কাছাকাছি। এই জনসংখ্যার বিচারে কলকাতা ভারতের চতুর্থ বৃহত্তম শহর ও তৃতীয় বৃহত্তম মেট্রোপলিটান বা মহানগরীয় অঞ্চল এবং বিশ্বের অষ্টম বৃহত্তম মহানগর অঞ্চল^[১] কলকাতা পৌরপ্রশাসকের উত্তর দিকে উত্তর চব্বিশ পরগনা, পূর্বে উত্তর ও দক্ষিণ চব্বিশ পরগনা এবং দক্ষিণ দিকে দক্ষিণ চব্বিশ পরগনা জেলা অবস্থিত। পশ্চিম দিকে হুগলী নদী এই শহরকে হাওড়া জেলা থেকে বিচ্ছিন্ন করেছে।

১৭৭২ সালে মুর্শিদাবাদ শহর থেকে বাংলার রাজধানী কলকাতায় স্থানান্তরিত করা হয়। ১৯১১ সাল পর্যন্ত কলকাতা শুধুমাত্র বাংলারই নয়, বরং সমগ্র ব্রিটিশ ভারতের রাজধানী ছিল। ১৯২৩ সালে ক্যালকটা মিউনিসিপ্যাল অথোরিটীর অধীনে কলকাতার স্থানীয় স্বায়তশাসন কর্তৃপক্ষ **কলকাতা পৌরসংস্থা** স্থাপিত হয়। ১৯৪৭ সালে ভারত বিভাগের পর কলকাতা নবগঠিত পশ্চিমবঙ্গ রাজ্যের রাজধানী ঘোষিত হয়। এই সময় কলকাতা ছিল আধুনিক ভারতের শিক্ষা, বিজ্ঞান, শিল্প, সংস্কৃতি ও রাজনীতির এক পীঠস্থান। ১৯৫৪ সালের পর থেকে রাজনৈতিক অস্থিরতা ও অর্থনৈতিক অবক্ষয়ের মূলে সেই পৌরব অনাকাঙ্ক্ষিত খর্ব হয়। তবে ২০০০ সালের পর থেকে এই শহর পুরনো অর্থনৈতিক ও বাণিজ্যিক সমৃদ্ধির পথে অগ্রসর হয় এবং সাংস্কৃতিক হজুয়ারব অনেকাংশেই পুনরুদ্ধার করে। যদিও ভারতের অন্যান্য শহরের মতো কলকাতাতেও নগরবেগনিত দারিদ্র্য ও পরিবেশ দূষণ একটি গুরুত্বপূর্ণ সমস্যা।

কলকাতা শহরের প্রসিদ্ধি এই শহরের বৈশ্বিক আন্দোলন ও সৃষ্টি সাংস্কৃতিক ঐতিহ্যের জন্য। ভারতের স্বাধীনতা আন্দোলন ও পরবর্তীকালে বামশক্তি গণআন্দোলনগুলিতে এই শহর এক বিশেষ ভূমিকা গ্রহণ করেছে। অন্যদিকে আধুনিক ভারতের প্রধান প্রধান সাংস্কৃতিক আন্দোলনগুলিরও প্রাণকেন্দ্র এই কলকাতা। এই কারণে এই শহরকে ভারতের *সাংস্কৃতিক রাজধানী* নামে অভিহিত করা হয়।^[১] অবার কলকাতা শহরে বিভিন্ন ভাষা, জাতি ও ধর্মাবলম্বী মানুষদের শান্তিপূর্ণ ও সৌহার্দ্যময় সহবাসের জন্য এই শহরকে *অনন্দ নগরী* বা *সিটি অফ অয়ে নামেও* অভিহিত করা হয়। রাজা রামমোহন রায়, রবীন্দ্রনাথ ঠাকুর, স্বামী বিবেকানন্দ, রোনাল্ড রস, সত্যেন্দ্রপ্রসাদ বসু, মাদার তেরেসা, সত্যজিৎ রায়, সত্যজিৎ রায়, সি চি রামন, অমর্ত্য সেন প্রমুখ বিশ্ববরেণ্য ব্যক্তিদের কর্মভূমি কলকাতা মহানগরী তার সমৃদ্ধ সাংস্কৃতিক ঐতিহ্যের জন্য আজও বিশ্ববাসীর চোখে মর্যাদার আসনে অধিষ্ঠিত।

কলকাতা

— **পশ্চিমবঙ্গের রাজধানী** —



Bengali Monolingual Retrieval

India Result Page - Windows Internet Explorer

http://www.clia.iitb.ac.in/clia-latest/result.jsp?TextArea=কলকাতা&start=10&hitsPerPage=10&sessionlang=bn&lang=en

AVG Total Protection AVG Info Get More Identity Guard

India Result Page

খোঁজ

খোঁজ : এই ইঞ্জিন দ্বারা অন্যান্য ইঞ্জিন দ্বারা

হিন্দি ইংরাজী বাংলা

পরিণাম 11 থেকে 20 মোট 35 খোঁজ --

[Calcutta Museums and Monuments,Indian Museum,Calcutta City Guide,Calcutta Travel Packages](#)

Calcutta Museums and Monuments,Indian Museum,Calcutta City Guide,Calcutta Fair/Festival,Calcutta Festival Tour,Online Booking of Calcutta Tour,Calcutta Travel A - Windows Internet Explorer

located in Chowringhee on the ...

কলকাতা মিউসিয়াম এবং স্মৃতিস্তম্ভ , ইন্ডিয়ান মিউজিয়াম , কোলকাতা শহর ...

করেছিলেন , সবচেয়ে পুরানো একটি এটা এশিয়া জাদুঘরের মধ্যে উপর টোরি ...

<http://www.calcutta-travel-packages.com/kolkata-museum.htm> (ভাঙার) (সারাংশ)

[calcutta Travel Packages,calcutta Sports,calcutta India Tours](#)

calcutta Travel Packages,calcutta Sports,calcutta India Tours and Travel,calcutta playing sports ...

কলকাতা ভ্রমণ প্যাকেজ , কলকাতা , কলকাতা , ভারতের ভ্রমণের এবং ভ্রমণ ...

<http://www.calcutta-travel-packages.com/kolkata-sports.htm> (ভাঙার) (সারাংশ)

[How to Reach,calcutta city guide,calcutta Travel Packages,ca](#)

How to Reach,calcutta city guide,calcutta Travel Packages,calcutta Transport cumbersome, ...

জন্য পৌঁছাতে হলে , কোলকাতা শহর নির্দেশিকা , কিভাবে কলকাতা ভ্রমণ প ...

পৌঁছাতে তথ্য তে সব কুশ্বসোমে হয় না করে থাকে , ...

<http://www.calcutta-travel-packages.com/reach.htm> (ভাঙার) (সারাংশ)

Calcutta Fair/Festival,Calcutta Festival Tour,Online Booking of Calcutta Tour,Calcutta Travel A - Windows Internet Explorer

http://www.calcutta-travel-packages.com/kolkata-festival.htm

AVG explore with YAHOO! SEARCH Total Protection AVG Info Get More Identity Guard

Kolkata festivals

Laxmi Puja
Time: October (5 days after Mahadashami)
Dedicated to: Laxmi, the Goddess of wealth, peace & prosperity Calcutta

Lakshmi puja festival falls in the month of October, about 5 days after Mahadashami. Laxmi Puja of Kolkata India is an important occasion, in which prayers are offered to Lakshmi, the Goddess of wealth, peace and prosperity. The puja of Ma Laxmi gives an opportunity to people to invite the Goddess of luck and prosperity to their homes. One or two days before the celebration of the festivity of Laxmi puja, the local markets of Kolkata are beautifully decorated. The shops get flooded with the idols of Laxmi seated on a lotus.

Saraswati Puja
Time: between late January and early February
Dedicated to: Saraswati, the Goddess of learning

Saraswati puja is one big occasion in Kolkata that takes place during the period between late January and early February. In fact, the day when puja of Ma Saraswati is done is declared as a state holiday. Calcutta Saraswati puja festival is celebrated with great pomp and show. It is on this day that the youngest female of the family is asked to dress up in yellow clothing. Sarswati puja of Kolkata India is dedicated to the goddess of learning.

Saraswati puja is conducted in almost every locality of Calcutta. People of the locality get together and assemble at the pandal to celebrate the festivity. Kids are really enthusiastic about the puja. It is during this puja that children pray to the goddess for their academic success. It is usually the pundit who performs the puja. After the puja is over, prasad is distributed to all.

Kali Puja

Kolkata festivals

- Tourist attractions
- Religious places
- People of Kalkata
- Entertainment in Kolkata
- Museums in Kolkata
- Foreign embassies
- Kalkata shopping guide
- Travel tips
- Kolkata sports
- Hotels
- Tour packages

Bengali –English Cross-lingual Retrieval

India Result Page - Windows Internet Explorer

http://www.clia.iitb.ac.in/cia-latest/result.jsp?TextArea=কলকাতা&sessionlang=bn&lang=hi&hitsPerPage=10

AVG Total Protection AVG Info Get More Identity Guard

Favorites New king of tech Apple o... Preschool Summer Them... Summer Coloring Pages 2... Printable Summer Colorin... Lecture Notes Suggested Sites Web Slice Gallery

India Result Page কলকাতা: W কলকাতা - উইকিপিডিয়া Best Time to Visit Calcu... Calcutta Fair/Festival, Ca...

INDIA search

কলকাতা

উন্নত খোঁজ
সাহায্য
কী-বোর্ড

খোঁজ

খোঁজ : এই ইঞ্জিন দ্বারা

হিন্দি ইংরাজী বাংলা

কলকাতা:
इसका पूर्व नाम अंग्रेजी में भले ही "कैलकटा" हो लेकिन बंगाल और बांग्ला में इसे हमेशा से कोलकाता एर पूर्व नाम इंग्लेज साले यदिओ येटा कैलकता' श्ये किछु बांग्ला एर बांग्ला साले परिचित छिल ...
<http://hi.pandapedia.com/wiki/कलकता> (आउर) (सारांश)

Dhapa »कलकता दया केन्द्र
Dhapa कलकता दया केन्द्र विकल्प: लाल हरा घर के बारे में भेजे एक पोस्ट Dhapa कोलकाता में धापा कलकता दया केन्द्र ता: लाल सबुज वाडी सम्पर्के भोजन एकाटि (पोखु Dhapa
<http://www.dhapa.com/hi/calcutta-mercy-centre> (आउर) (सारांश)

मौसम कलकता में | भारत - Notepad, संस्कृति, स्मारको और तस्वीरें
भारत में दस शहरों के लिए मौसम का पूर्वानुमान और अच्छी छुट्टी सेवा मतलब मई:) यह कलकता भारतेर १० शहरेर जन्य जलवायु उपलक्षण पूर्वानुमान एर छुट्टिेर परिषेवा अर्थ (म:)
<http://india.inet.net/2009/01/15/season-in-kolkata/> (आउर) (सारांश)

कलकता: - Windows Internet Explorer

http://hi.pandapedia.com/wiki/%E0%A4%B2%E0%A4%B5%E0%A4%B8%E0%A4%BD%E0%A4%BE

AVG Total Protection AVG Info Get More Identity Guard

Favorites New king of tech Apple o... Preschool Summer Them... Summer Coloring Pages 2... Printable Summer Colorin... Lecture Notes Suggested Sites Web Slice Gallery

कलकता: कलकता: W कलकता - উইকিপিডিয়া Best Time to Visit Calcu... Calcutta Fair/Festival, Ca...

आधिकारिक जालस्थल: www.kolkatamcity.com

Footnotes

निर्देशांक: 22°34'11"N 88°22'11"E / 22.5697, 88.3697 बंगाल की खाड़ी के शीर्ष तट से १८० किलोमीटर दूर हुगली नदी के बायें किनारे पर स्थित **कोलकाता (बांग्ला):** नाम कलकता पश्चिम बंगाल की राजधानी है। यह भारत का दूसरा सबसे बड़ा महानगर तथा पाँचवा सबसे बड़ा बन्दरगाह है। यहाँ की जनसंख्या १,३२,१६,५४६ है।^[1] इस शहर इतिहास अत्यंत प्राचीन है। इसके आधुनिक स्वरूप का विकास अंग्रेजों एवं फ्रांस के उपनिवेशवाद के इतिहास से जुड़ा है। आज का कोलकाता आधुनिक भारत के इतिहास की गाथाएँ अपने आप में समेटे हुये है। शहर को जहाँ भारत के शैक्षिक एवं सांस्कृतिक परिवर्तनों के प्रारम्भिक केन्द्र बिन्दु के रूप में पहचान मिली है वहीं दूसरी ओर इसे भारत साम्यवाद आंदोलन के गढ़ के रूप में भी मान्यता प्राप्त है। महलों के इस शहर को सिटी ऑफ जाय के नाम से भी जाना जाता है।

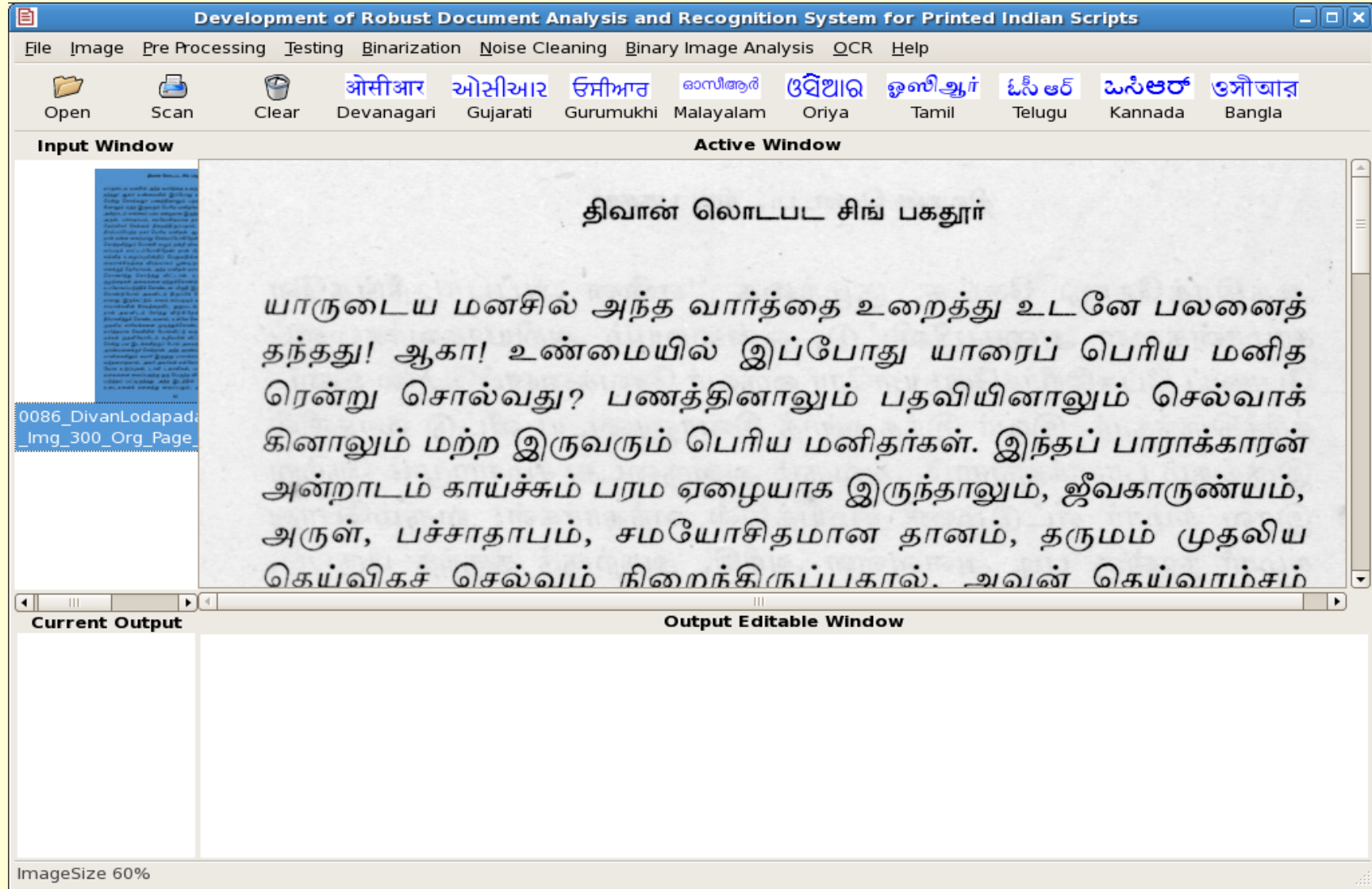
अपनी उत्तम स्थिति के कारण कोलकाता को पूर्वी भारत का प्रवेश द्वार कहा जाता है। यह रेलमार्ग, वायुमार्ग तथा सड़क मार्गों द्वारा देश के विभिन्न भागों से जुड़ा हुआ है। यह यातायात का केन्द्र, विस्तृत बाजार वितरण केन्द्र, शिक्षा केन्द्र, औद्योगिक केन्द्र तथा व्यापार का केन्द्र है। अजायबघर, चिड़ियाखाना, बिरला तारमंडल, हावड़ा पुल, कालीघाट विलियम, विक्टोरिया मेमोरियल, विज्ञान नगरी आदि मुख्य दर्शनीय स्थान हैं। कोलकाता के निकट हुगली नदी के दोनों किनारों पर भारतवर्ष के प्रायः अधिकांश जूट के कार अवस्थित हैं। इसके अलावा मोटरगाड़ी तैयार करने का कारखाना, सूती-वस्त्र उद्योग, कागज-उद्योग, विभिन्न प्रकार के इंजीनियरिंग उद्योग, जूता तैयार करने का कारखाना, हथियार उद्योग एवं चाय विक्रय केन्द्र आदि अवस्थित हैं। पूर्वांचल एवं सम्पूर्ण भारतवर्ष का प्रमुख वाणिज्यिक केन्द्र के रूप में कोलकाता का महत्त्व अधिक है।

अनुक्रम

- १ विकास और नामकरण
- २ इतिहास
 - २.१ स्वाधीनता आंदोलन में भूमिका
 - २.२ बाबू संस्कृति और बंगाली पुनर्जागरण
 - २.३ आधुनिक कोलकाता
 - २.४ अर्थ व्यवस्था

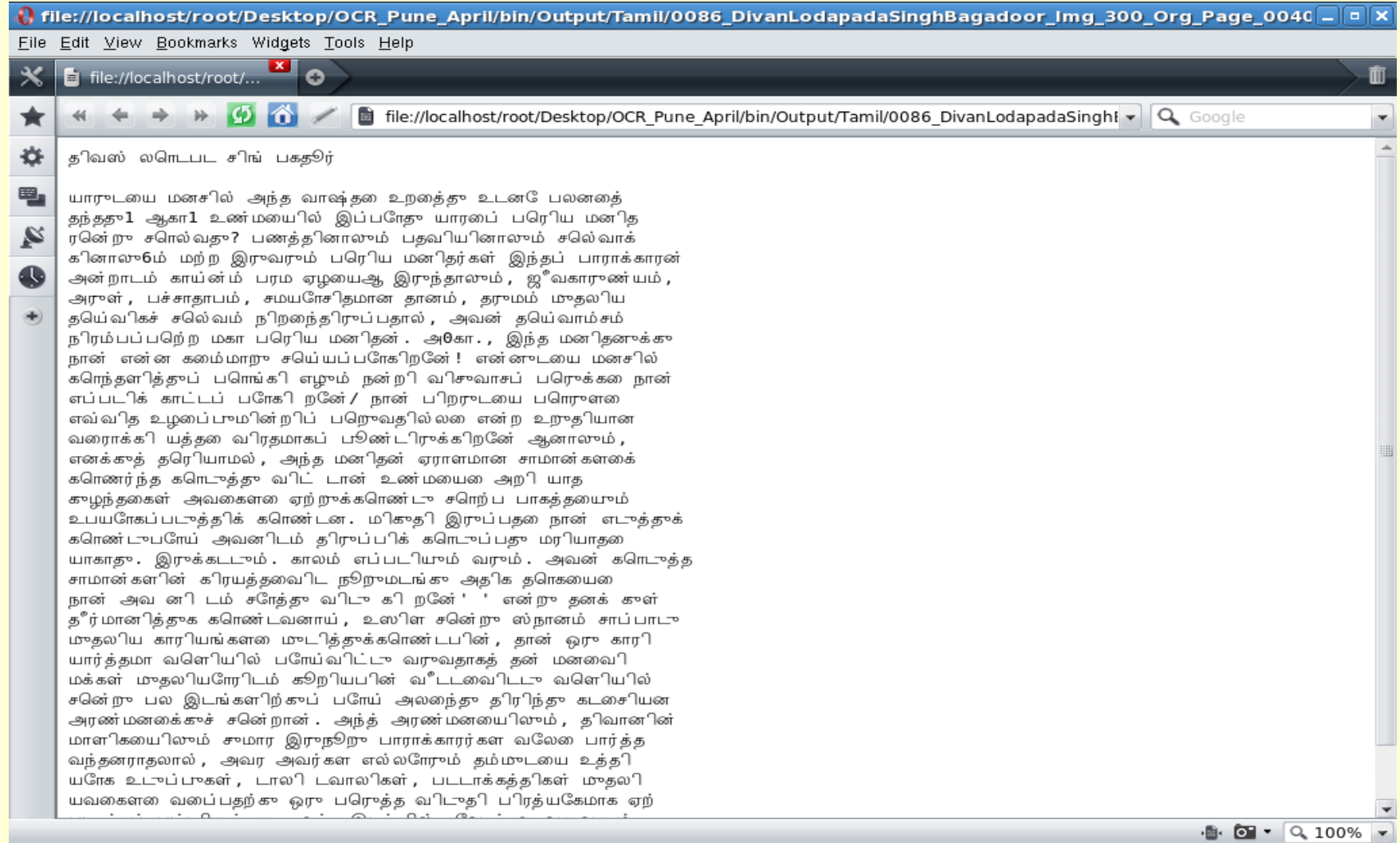
Bengali –Hindi Cross-lingual Retrieval

OCR



Original Image Loaded in Input Window of OCR

OCR



Output of Tamil OCR

Sample Tamil OHWR form

D:\Working Versions\Online ICRWGeneral\Inspection Report\Traffic Inspection Report(Application) For Unicode Common [(15-10-09)\top\files 25-0...

File View Recognize Edit Language Help

X-Shift 1 X-Scale 0 Y-Shift 1 Y-Scale 0 C50.top 1

வினாக்கள் உள்ளனவகையில் வினாக்கள் அல்லாதவகையில் அல்லாதவகையில் அல்லாதவகையில்

பயிற்சி எண்: 2748

1. பற்றி எண்: 2748

2. பற்றி எண்: 2748

3. பற்றி எண்: 2748

4. பற்றி எண்: 2748

5. பற்றி எண்: 2748

6. பற்றி எண்: 2748

7. பற்றி எண்: 2748

8. பற்றி எண்: 2748

9. பற்றி எண்: 2748

10. பற்றி எண்: 2748

11. பற்றி எண்: 2748

12. பற்றி எண்: 2748

13. பற்றி எண்: 2748

14. பற்றி எண்: 2748

15. பற்றி எண்: 2748

16. பற்றி எண்: 2748

17. பற்றி எண்: 2748

18. பற்றி எண்: 2748

19. பற்றி எண்: 2748

20. பற்றி எண்: 2748

21. பற்றி எண்: 2748

22. பற்றி எண்: 2748

23. பற்றி எண்: 2748

24. பற்றி எண்: 2748

25. பற்றி எண்: 2748

26. பற்றி எண்: 2748

27. பற்றி எண்: 2748

28. பற்றி எண்: 2748

29. பற்றி எண்: 2748

30. பற்றி எண்: 2748

31. பற்றி எண்: 2748

32. பற்றி எண்: 2748

33. பற்றி எண்: 2748

34. பற்றி எண்: 2748

35. பற்றி எண்: 2748

36. பற்றி எண்: 2748

37. பற்றி எண்: 2748

38. பற்றி எண்: 2748

39. பற்றி எண்: 2748

40. பற்றி எண்: 2748

41. பற்றி எண்: 2748

42. பற்றி எண்: 2748

43. பற்றி எண்: 2748

44. பற்றி எண்: 2748

45. பற்றி எண்: 2748

46. பற்றி எண்: 2748

47. பற்றி எண்: 2748

48. பற்றி எண்: 2748

49. பற்றி எண்: 2748

50. பற்றி எண்: 2748

Technology development for inclusive growth

Text to Speech in Indian Languages

To develop accessible technologies for differently abled section of the society , TDIL **programme** has undertaken initiatives.

Consortium mode project has been initiated for development of:

Text to Speech System with Braille Interface in six Indian Languages : Hindi , Tamil , Telugu , Bengali , Marathi and Malayalam Languages

Consortium Leader : IIT Madras

Consortium Members : IIT Kharagpur

IIT Guwahati

IIIT Hyderabad

C-DAC Mumbai

C-DAC Thiruvananthapuram

Localized TDIL data Centre Portal in Mozilla Firefox

TDIL Data Center - Mozilla Firefox

फ़ाइल (E) संपादन (E) अवलोकन (V) जाओ (G) पृष्ठसंकेत (B) औज़ार (T) सहायता (H)

नयी खिड़की (N) Ctrl+N
नया टैब (T) Ctrl+T
स्थान खोलें... (L) Ctrl+L
फाइल को खोलें... (O) Ctrl+O
बन्द करें (C) Ctrl+W

पृष्ठ को संचित करें जैसे... (A) Ctrl+S
कड़ी भेजें... (E)

पृष्ठ सेटप... (U)
छपाई पृष्ठ का दृश्य (V)
छपाई(P)... Ctrl+P

आयात... (I)
ऑफ़लैन कार्य (W)
निर्गम (X)

http://ildc.in/Hindi/Hindex.aspx

Help

सत्यमेव जयते

भारतीय भाषाओं के लिए प्रौद्योगिकी विकास
Technology Development for Indian Languages

मुख्य पृष्ठ | भाषा विकल्प | डाउनलोड | सामान्य प्रश्न | Help Manual | संपर्क करें | प्रतिपुष्टि | Site Map

पृष्ठभूमि
पंजीकरण
सीडी का अनुरोध

यदि हिंदी में नहीं देख पा रहे हैं तो यहाँ क्लिक करें

जन जन को एक सूत्र में पिरोती भारतीय भाषा प्रौद्योगिकी

एक अरब से भी अधिक बहुभाषी भारतवासियों को परस्पर समीप लाने में सूचना प्रौद्योगिकी की भाषा तकनीकी एक अहम् भूमिका निभाती है।

भाषा तकनीकी में विकसित उपकरणों को जनसामान्य तक पहुँचाने हेतु भारत सरकार के सूचना प्रौद्योगिकी विभाग के प्रावधान के अंतर्गत www.ildc.gov.in तथा www.ildc.in वेबसाइटों के द्वारा व्यवस्था की गई है।

इन उपकरणों एवं सेवाओं में मुख्य हैं-

फ्रॉन्ट	कोड परिवर्तक	वर्तनी संशोधक	ओपन ऑफिस
गैंगेन्ना	ई-गेन कन्वर्टर	ओपी आर	शट्टकरोण

News and Events

Tamil Software Tools CD 2010
CD CD CD CD CD CD CD CD CD CD

लॉगइन

लॉगइन

पासवर्ड

पूरा हुआ

start TDIL Data Center - M... EN 11:44 AM

Localized TDIL data Centre Portal in Mozilla Firefox

TDIL Data Center - Mozilla Firefox

फ़ाइल (F) संपादन (E) अवलोकन (V) जाओ (G) पृष्ठसंकेत (B) ऑज़र (I) सहायता (H)

ऑज़र पट्टी (I)
 ✓ वस्तुस्थिति पट्टी (B)
 किनारे की पट्टी (E)
 रोको (S) Esc
 पुनः लोड करें (R) Ctrl+R
 सूत्र परिमाण (S)
 पृष्ठ शैलि (Y)
 अक्षर ऐनकोडिंग (C)
 पृष्ठ स्रोत (G) Ctrl+U
 सम्पूर्ण पढ़ा (E) F11

/Hindi/Hindex.aspx

जायें

Select Keyboard

टी डी एल
 अरुअरअरअरअरअरअरअरअरअर

भारतीय भाषाओं के लिए प्रौद्योगिकी विकास
 Technology Development for Indian Languages

सत्यमेव जयते

मुख्य पृष्ठ | भाषा विकल्प | डाउनलोड | सामान्य प्रश्न | Help Manual | संपर्क करें | प्रतिपुष्टि | Site Map

पृष्ठभूमि
 पंजीकरण
 सीडी का अनुरोध

यदि हिंदी में नहीं देख पा रहे हैं तो यहाँ क्लिक करें

जन जन को एक सूत्र में पिरोती भारतीय भाषा प्रौद्योगिकी

एक अरब से भी अधिक बहुभाषी भारतवासियों को परस्पर समीप लाने में सूचना प्रौद्योगिकी की भाषा तकनीकी एक अहम् भूमिका निभाती है।

भाषा तकनीकी में विकसित उपकरणों को जनसामान्य तक पहुँचाने हेतु भारत सरकार के सूचना प्रौद्योगिकी विभाग के प्रावधान के अंतर्गत www.ildc.gov.in तथा www.ildc.in वेबसाइटों के द्वारा व्यवस्था की गई है।

इन उपकरणों एवं सेवाओं में मुख्य हैं-

फ्रॉन्ट	कोड परिवर्तक	वर्तनी संशोधक	ओपन ऑफिस
गैमिंग	ई गेन क्लानांग	ओ गी अंग	शट्टकोश

News and Events

NEW !!
 NEW !!

लॉगइन

लॉगइन

पापवर्ड

पूरा हुआ

start

TDIL Data Center - M...

1 - Paint

EN

11:46 AM

Linguistic Resources Developed under TDIL Programme

Written Text Resources

- **Parallel Corpora:** One Million pages Parallel Corpora with graphical user interface in 13 languages namely English, Hindi, Punjabi, Tamil, Telugu, Kannada, Malayalam, Bengali, Oriya, Marathi, Assamese, Gujarati and Nepali languages
- **Bi-lingual Dictionaries** Bi-lingual Dictionaries of English-Hindi, English-Bengali, English-Telugu, English-Tamil, English-Kannada, English-Malayalam, English-Oriya and Urdu-Hindi, each with over 30,000 root words
- **Ontology & Word-Net:** Hindi Word-net with 30000 sync-sets with morphological analyzer and front-end. Oriya Word-net with 1100 lexical entries with X-window interface.
- **On-line Vishwakosh** (Encyclopaedia in Hindi) with 9162 topics
- **Phrasal Dictionaries** in Tamil and Kannada languages
- **Information Technology Terminology (10000 terms)** in Hindi
- **Text corpora** of 3 Million words in major Indian languages .

Speech Resources:

Speech Corpora:

- Annotated Speech Corpora of approximately 50 hours has been developed for 10 Indian Languages namely Hindi, Marathi, Punjabi, Bengali, Assamese , Manipuri. Tamil, Malayalam, Telugu and Kannada languages
- The Speech Corpora with sample sound are available at Indian Language Data Centre (<http://tdil-dc.in/>)

Semi-Automatic Annotation Tool for Speech Corpora

- Semi-automatic annotation tool for speech database has also been developed
- Five levels of annotation namely phoneme, syllable, word, phrase and parts of speech (POS) are used for annotation.
- The Annotated speech signal and its output i.e. standard format time table (SFT) are available.

Linguistic resources development under Consortium Mode projects

- Multilingual Sense Dictionary in 6 Indian Languages Pairs
- UNL based Information Extraction modules
- Morphological Analyzers for major Indian Languages
- Font Transcoder
- Indian Languages Annotated Corpus for Tourism and Health Domains for 11 Indian Languages
- Word-net for 4 Indian Languages Assamese, Bodo , Manipuri and Nepali languages.
- Speech Corpus in 6 Indian Languages Hindi, Bengali, Tamil, Telugu, Malayalam and Marathi languages for Text-to-speech system applications.

Hindi Word-Net

Hindi WordNet - Windows Internet Explorer
http://www.cfil.t.iitb.ac.in/wordnet/webhwn/wn.php

File Edit View Favorites Tools Help

Indian Language Technolo... hindi wordnet - Google Se... Hindi WordNet

हिन्दी शब्दतंत्र Hindi Wordnet
हिंदी शब्द संकल्पनाकोश A Lexical Database for Hindi

Total unique words:84346 | Total Synsets: 34418
Linked Synsets: 15062 | Last Updated: 10 Nov 2010

हिन्दी शब्द खोजें (Write in Devanagari):
 खोजें(Search)

Devanagari Keyboard

उदाहरण Examples सहायता Help प्रतिक्रिया दें Give feedback प्रतिक्रियाएँ देखें Previous feedbacks लॉग देखिए View log

पुराना इंटरफेस Previous interface मराठी शाब्दबंध Marathi Wordnet हिन्दी-अंग्रेजी शब्दकोश Hindi-English Dictionary

सी.एफ.आई.एल.टी. मुख्यपृष्ठ CFILT home हिन्दी शब्दतंत्र Hindi Wordnet

English data courtesy Wordnet 2.1
CFILT, IIT Bombay.

4 Internet Explorer Research Collaborat... wai mails - Notepad 4 Microsoft Office... EN 15:09

Oriya Word-Net

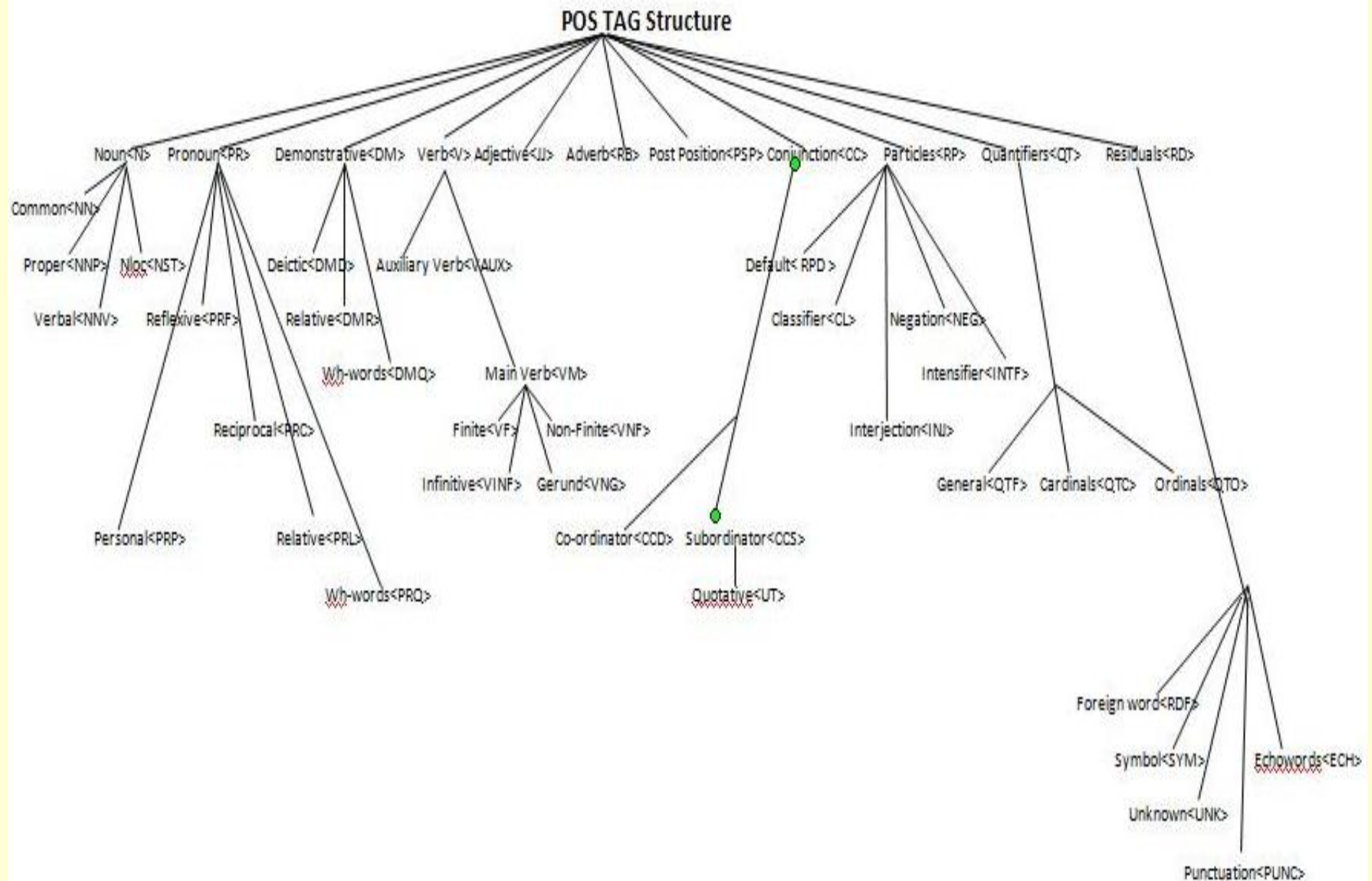
The screenshot displays the 'WORDNET FOR ORIYA' application window. The title bar includes standard window controls. The main interface features a menu bar with 'File', 'History', and 'Help'. Below this is a search bar with the text 'SEARCH FOR WORD --' and a text input field containing 'bhaLa'. To the right of the search bar is a small icon. Below the search bar, there are tabs for 'Noun' and 'Adjective', with 'Noun' currently selected. A dropdown menu is open over the 'Noun' tab, listing the following options: 'Synonymy', 'Antonymy', 'Hypernyms(..is a kind of..)', 'Meronymy', and 'Morphology'. The main content area shows the search results for 'bhaLa' as an adjective. It lists three senses: (1) 'bhaLa' (good, au...), (2) 'bhaLa' (ଅତି ଭଲ), and (3) 'bhaLa' (aphorisms for solution of arithmetical calcution composed by subhamkari) ଅନୁକୂଳ. To the right of the main content area, there is a vertical scroll bar and a small text area containing the text '...lthy) used in following sense in Orinet--' and 'ଓ ଆନନ୍ଦପାଠକ'. At the bottom of the window, there is a cyan-colored bar with the text 'Overview of word ଭଲ'.

STANDARDIZATION EFFORTS

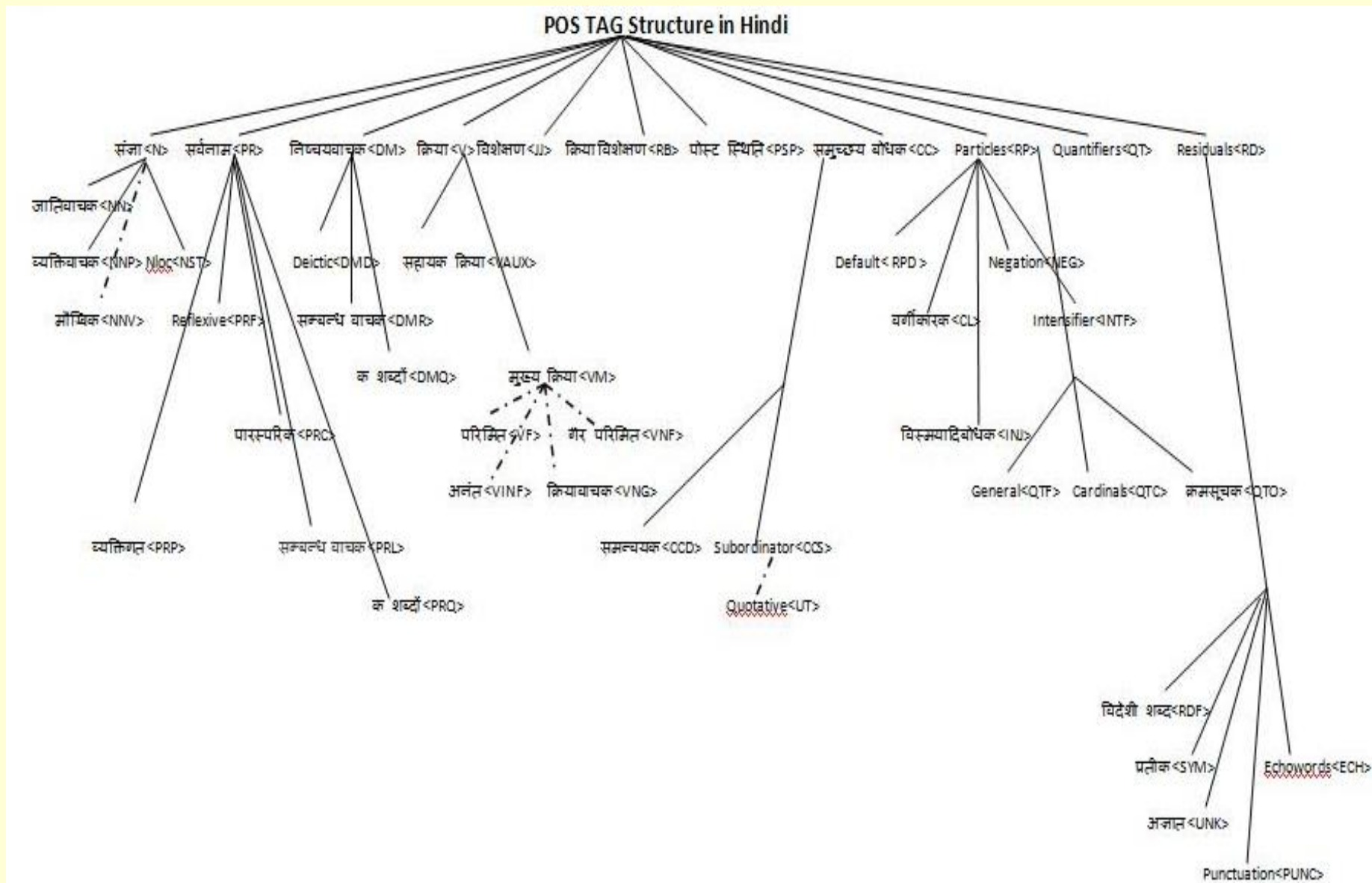
STANDARDIZATION EFFORTS

- **UNICODE:** Department of Information Technology is the voting member of the Unicode Consortium to ensure the adequate representation of Indic scripts in the Unicode Standards.
- **Common Locale Data Repository:** Modifications/ Development of UNICODE Common Locale data repository (CLDR) containing major fields dates, times, time zones, numbers, and currency values; sorting text; etc.
- **Language Tags:** The Language Tag Standard ISO 639-x (x stands for different versions) are being used in many other international Standards and Best Practices such as IETF (Internet Engineering Task Force) RFC 4646, RFC 4647 and W3C web standards.
- **Web Standards (W3C):** Major Initiative has also been undertaken for adequate representation of Indian Language Specificities in W3C existing and futuristic web standards. W3C India Office has been setup in Department of IT, New Delhi.
- **Script Grammar:** The nonlinear nature of Indian Scripts requires standardization of Script Grammars.
- Initiatives have also been undertaken for standardization of Domain Names in Indian Languages and IPA representation for Indian Languages.

Part Of Speech Tagging



Part Of Speech structure in Hindi



⊗ ... - - Dotted line shows tags are not included in Hindi POS

Features of XML Schema

- It is easier to describe allowable document content
- It is easier to validate the correctness of data
- It is easier to work with data from a database
- It is easier to define data facets (restrictions on data)
- It is easier to define data patterns (data formats)
- It is easier to convert data between different data types

XML Schema for POS tag -Hindi

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <file Desc>
  <titleStmt>
  <title>POS tag in hindi</title>
  <script>devnagari</script>
  <language>hindi</language>
    <label language>.....</label language>
  <type>bimodal</type>
  <xs:element name="cat" POS cat="noun" hcat="संज्ञा"tag="N">
  <xs:attribute name="type" subcat="common" hcat="जातिवाचक" tag="NN" />
  <xs:attribute name="type" subcat="Proper" hcat="व्यक्तिवाचक" tag="NNP" />
  <xs:attribute name="type" subcat="Verbal" hcat="मौखिक" tag="NNV" />
  <xs:attribute name="type" subcat="Nloc" hcat="" tag="NST" />
  </xs:element></xs:schema>
```

TTS Corpus specification

1	Recording Instrument (Lab. Environment)	Dynamic Mic. With frequency response 80Hz-20 kHz (equivalent to Shure, Sennheiser, etc.) Preamp.: 30Hz-15kHz Sound Card: Creative Gold
2	Recording Environment	Speech studio (SNR \geq -45 dB)
3	Recording Format	16bit PCM Mono, 48.0 kHz
4	Informant Selection	Standard ITU-T (Annexure-1) , Age should be 25-35.
	Speech rate	Medium
	Emotion	Neutral
	Style	Read out
5	No. of Informant	2(1Male & 1Female)
6	Contents	Sentences →Cover all the di-phone, syllable and most probable tri-phone at least 2 occurrence probability. About 1000 phonetically balanced (PB)sentences. Paragraph (at least 5 sentences): 10-20 which more or less covers different prosody variation (Desirable: 3 repetitions of same data) Story : 2-3 stories of 4-5 paragraphs.
7	Annotation Hierarchy	Acoustic →Phone, Syllable, Word <i>Note : the definition of the Phone, Syllable, Word boundary in continuous</i> <i>Speech as given in annotation guidele (Annexure-II)</i> Linguistic → POS (Functional), Phrase, Clause

Testing & Evaluation

Machine Translation : Indian Language to Indian Language

Guideline for Evaluation : on 5 point scale

5. Perfect : (like some one who knows the language)
4. Comprehensible, occasional errors : (like some one speaking Hindi getting all its genders wrong)
3. Comprehensible but has quite a few errors : (like some one who can speak your language but would make lots of error. However, you can make sense out of what is being said.)
2. Some parts make sense but is not comprehensible over all : (like listening to a language which has lot of borrowed words from your language- you understand those words but nothing more)
1. Non-Sense : (if the sentence doesn't make any sense at all – It is like some one speaking to you in a language you don't know)

Machine Translation : Indian Language to Indian Language

System Level Performance of randomly chosen web text of ILMT Project

S.No.	Systems	Comprehensibility (in %) (with score 3-5]	Marginal comprehensibility (in %) (with score 2.6 - 5]	Promised comprehensibility (in%) (at the end of phase I)
1	Telugu-Tamil	97.60	99.00	97.60
2	Punjabi-Hindi	93.00	96.00	95.00
3	Urdu-Hindi	84.00	89.30	85.00
4	Hindi-Punjabi	77.40	88.00	80.00

Machine Translation : Indian Language to Indian Language

System Level Performance of randomly chosen web text of ILMT Project

S.No	Systems	Comprehensibility (in %) (with score 3-5]	Marginal comprehensibility (in %) (with score 2.6 - 5]	Expected comprehensibility (in%) (at the end of phase I)
5	Hindi- Telugu	42.00	67.42	60.00
6	Hindi- Urdu	56.00	66.60	60.00
7	Bengali- Hindi	24.00	46.60	50.00
8	Hindi- Bengali	10.60*	26.60	50.00
9	Tamil- Hindi	36.00#	63.00#	50.00

* : Debugging of linguistic integration is going on

: Performance on Tourism domain

CLIA Testing

Salient points of the strategy:

- access to data
- collecting the output for different categories of input
- testing modalities, training workshop, grading
- data analysis

IMPLEMENTATION

- Deployment of final CLIA system
- Testing and culling of data:
 - Snippet Generation
 - Snippet Translation
- Evaluation
 - Identification of evaluators
 - Workshop to train the evaluators for all the six different languages

Grading Scales

Snippet Translation:

0	The translated output is not at all comprehensible to you.
1	Keywords comprehensible and rest of the snippet makes little sense.
2	Comprehensible after accessing the source (English/Hindi) text.
3	Comprehensible with difficulty.
4	Snippet is comprehensible.

Snippet Generation:



0	Snippet judged totally divergent from the page
1	Snippet is vague and does not provide a very clear idea of the page
2	The Snippet allows the user to understand the page but is not still very clear
3	The Snippet provides a clear picture of the page but its grammatical accuracy is low.
4	The Snippet is consistent with the page and is couched in correct language

Future directions

- The complexity and vastness of Indian Language Ecosystem requires sustained and collaborative efforts for development and standardization of Linguistic Resources and Tools towards development of ICT solutions in Indian Languages.
- Comprehensive policy for standardization, testing and evaluation of Linguistic Resources and Tools are being planned.
- Challenges in replication of the development of linguistic resources and tools in all 22 Indian languages.
- **Testing and Evaluation campaigns inline of those of international efforts like CLEF , NIST and Blizzard Challenge are also being initiated.**

