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

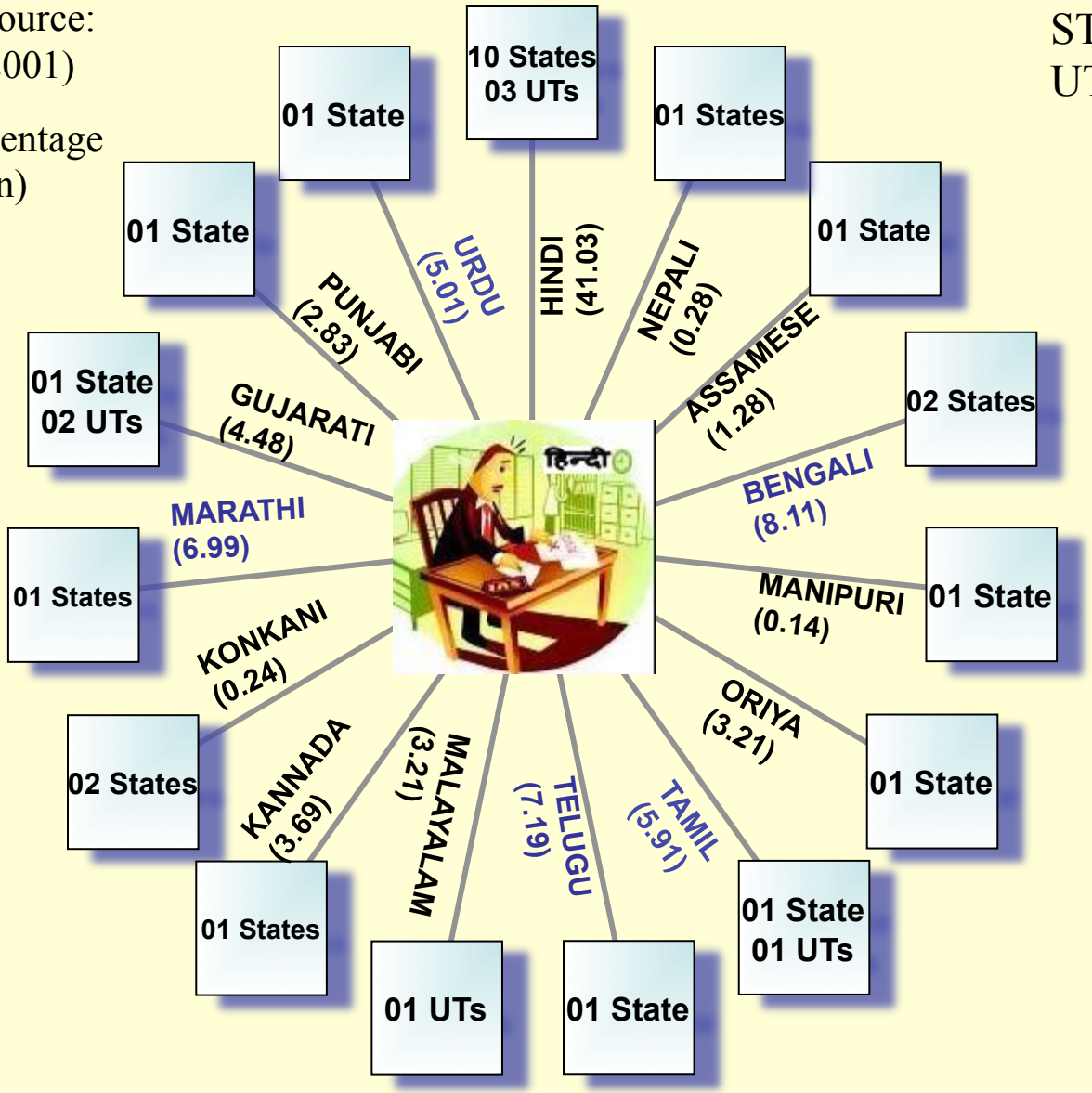
Organization of presentation:

- 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)

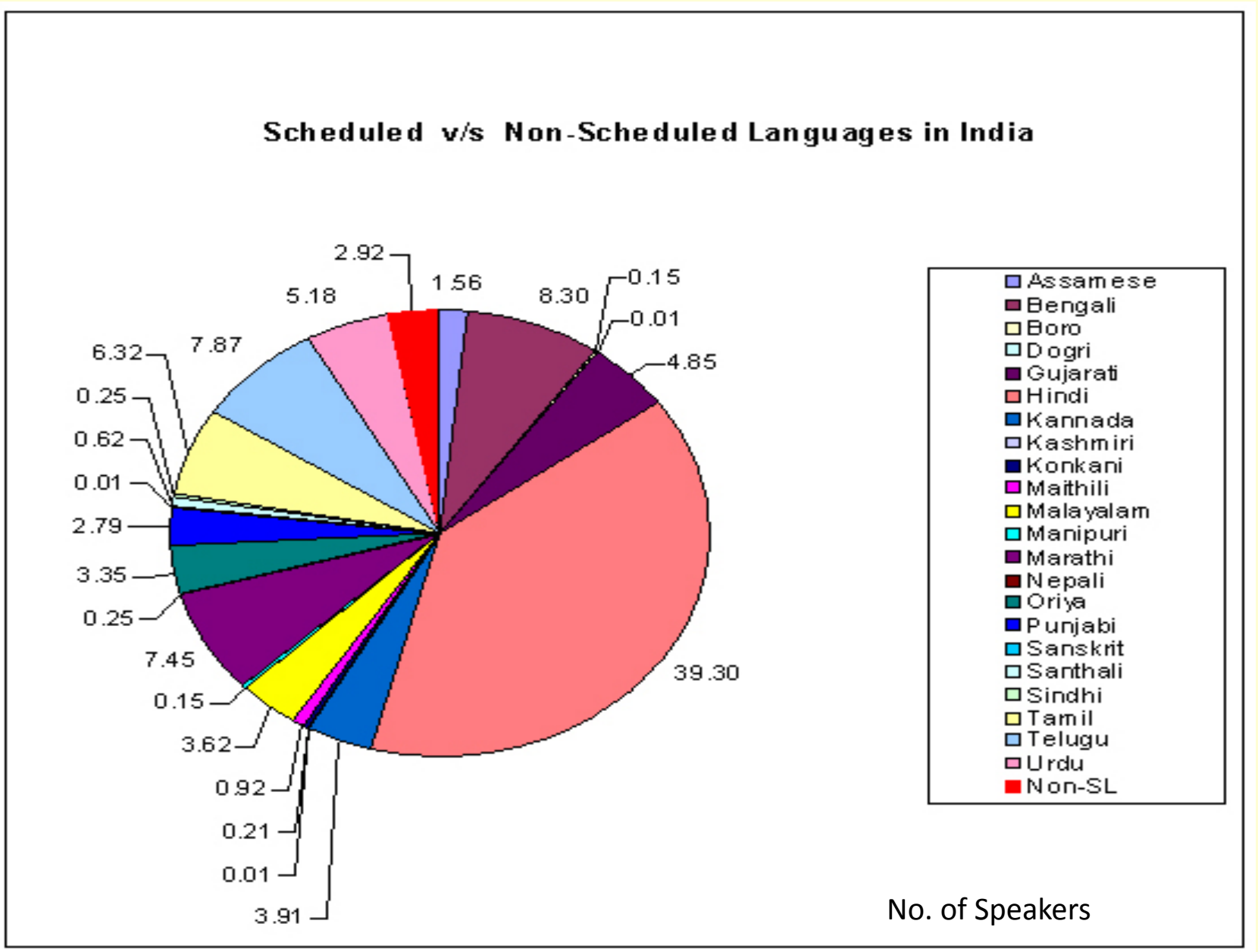
INDIA
STATES: 28
UT: 07



Linguistic Scenario in India

Source – Census 2001, India

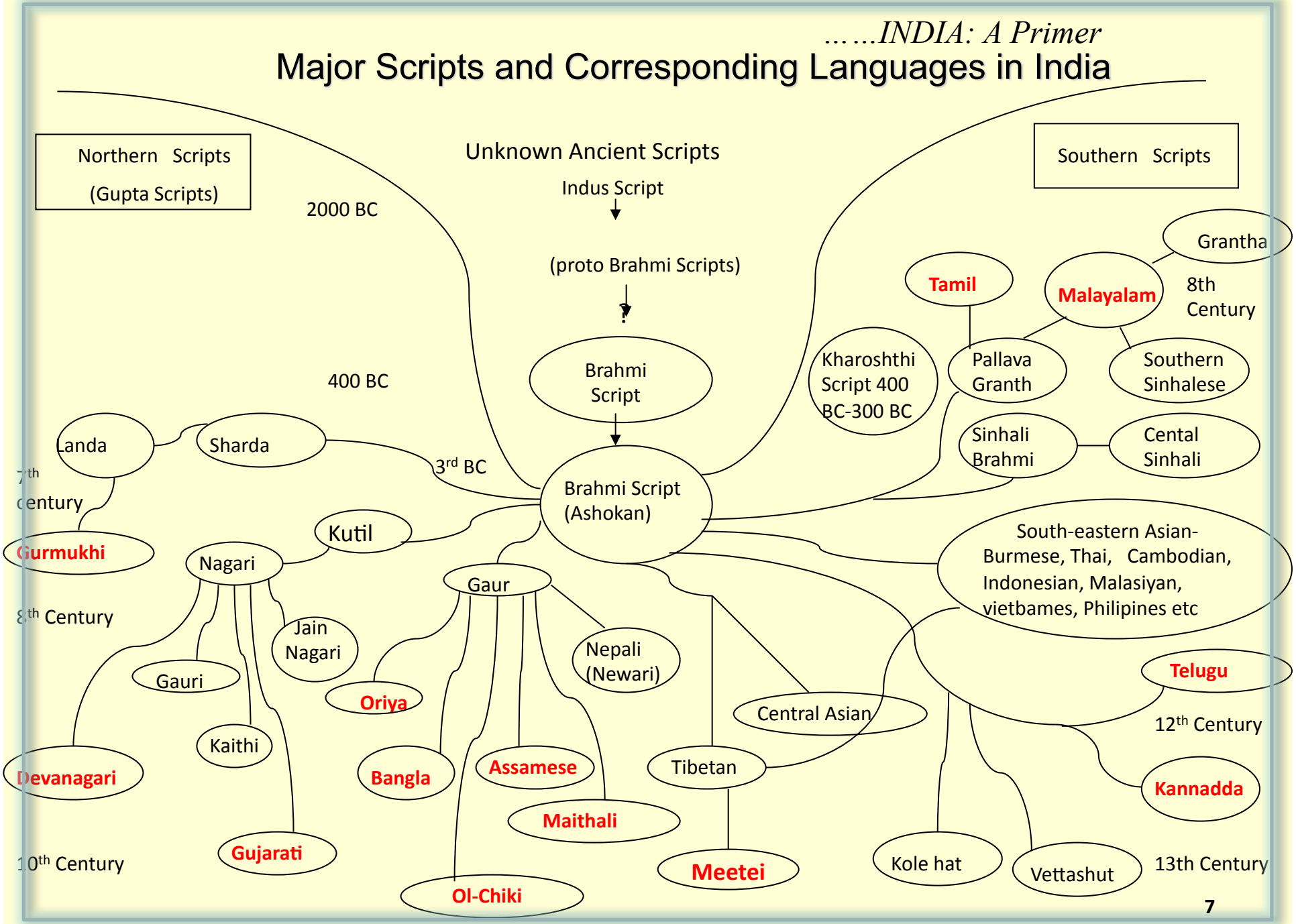
Language	Speakers	Percentage to total population	State(s)
Assamese	13,168,484	1.28	Assam
Bengali	83,369,769	8.11	Andaman & Nicobar Islands, Assam, Tripura, West Bengal
Bodo	1,350,478	0.13	Assam
Dogri	2,282,589	0.22	Jammu and Kashmir
Gujarati	46,091,617	4.48	Dadra and Nagar Haveli, Daman and Diu, Gujarat
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
2.92
1.56
8.30
0.15
0.01
4.85
5.18
7.87
6.32
0.25
0.62
0.01
2.79
3.35
0.25
7.45
0.15
3.62
0.92
0.21
0.01
3.91
39.30
Assamese
Bengali
Boro
Dogri
Gujarati
Hindi
Kannada
Kashmiri
Konkani
Maithili
Malayalam
Manipuri
Marathi
Nepali
Oriya
Punjabi
Sanskrit
Santhali
Sindhi
Tamil
Telugu
Urdu
Non-SL
No. of Speakers

Official Indian Languages & Scripts

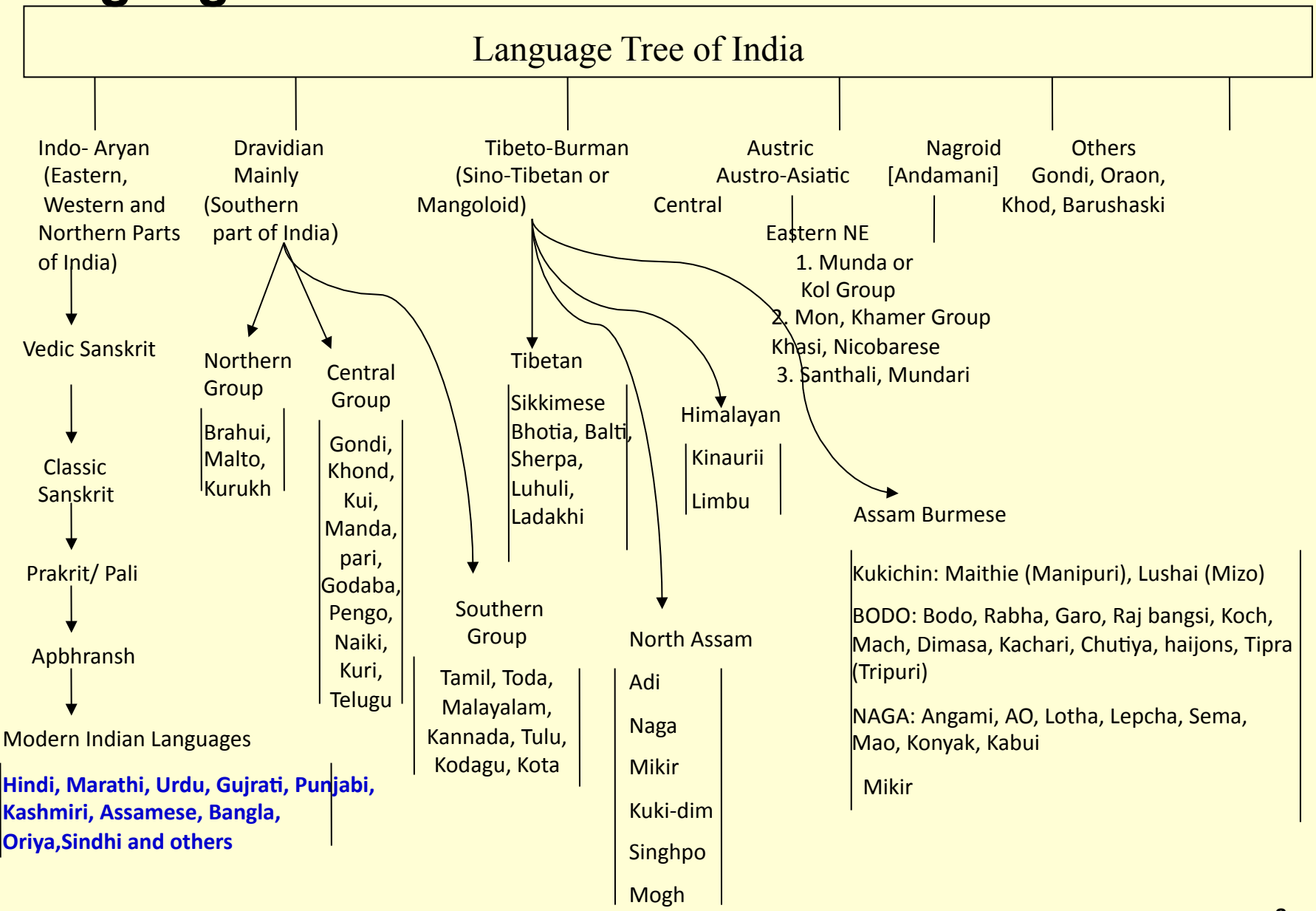
Sl. No.	Language	Script
1.	Hindi	Devanagari
2.	Sanskrit	Devanagari
3.	Marathi	Devanagari
4.	Konkani	Devanagari
5.	Nepali	Devanagari
6.	Maithili	Devanagari
7.	Sindhi	Devanagari
8.	Bodo	Devanagari
9.	Dogri	Devanagari
10.	Bengali	Bengali
11.	Assamese	Bengali
12.	Manipuri	Bengali, Meetei (Mayak)
13.	Gujarati	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

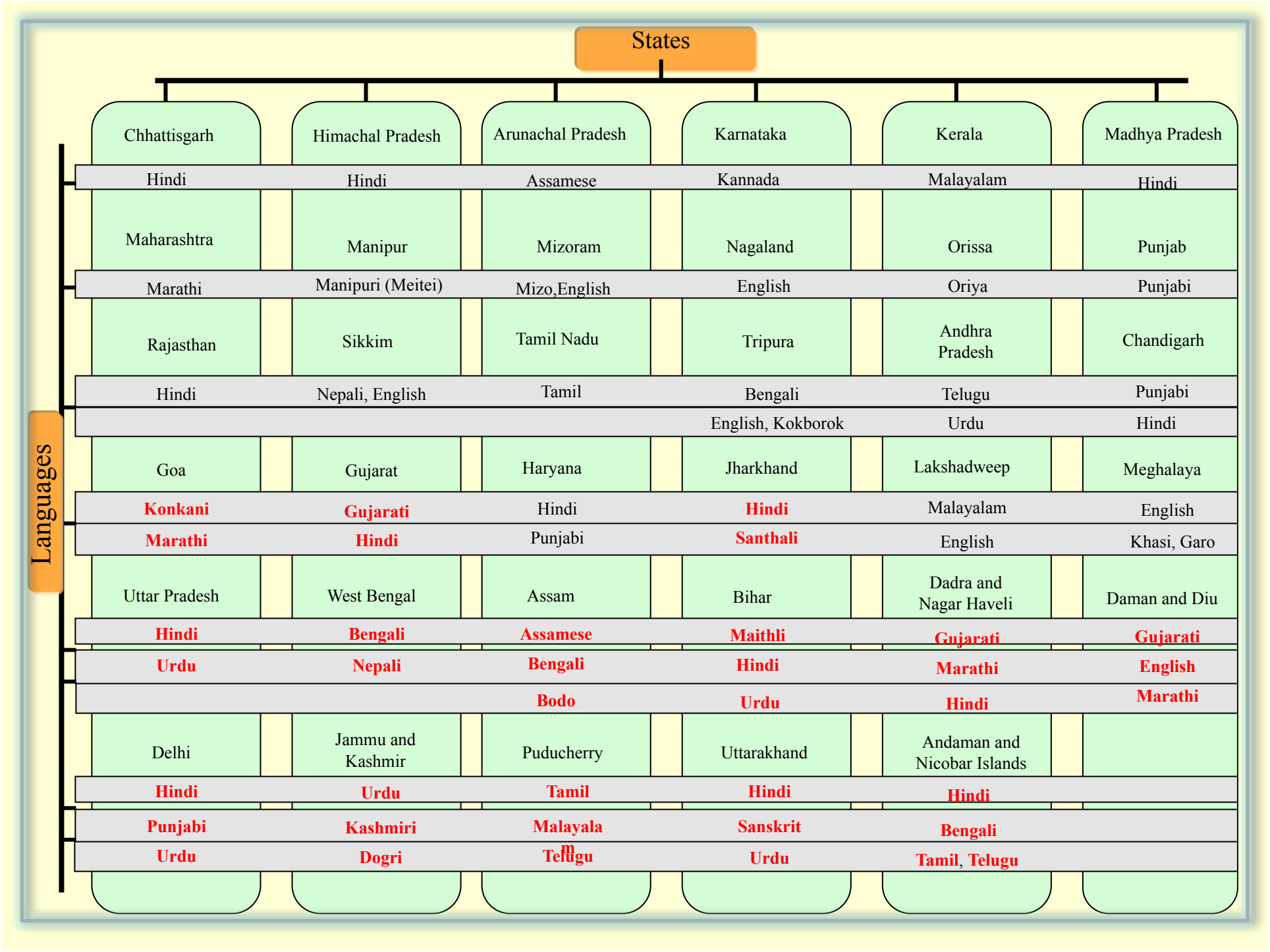
Major Scripts and Corresponding Languages in India



Languages of India

……INDIA: A Primer




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

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

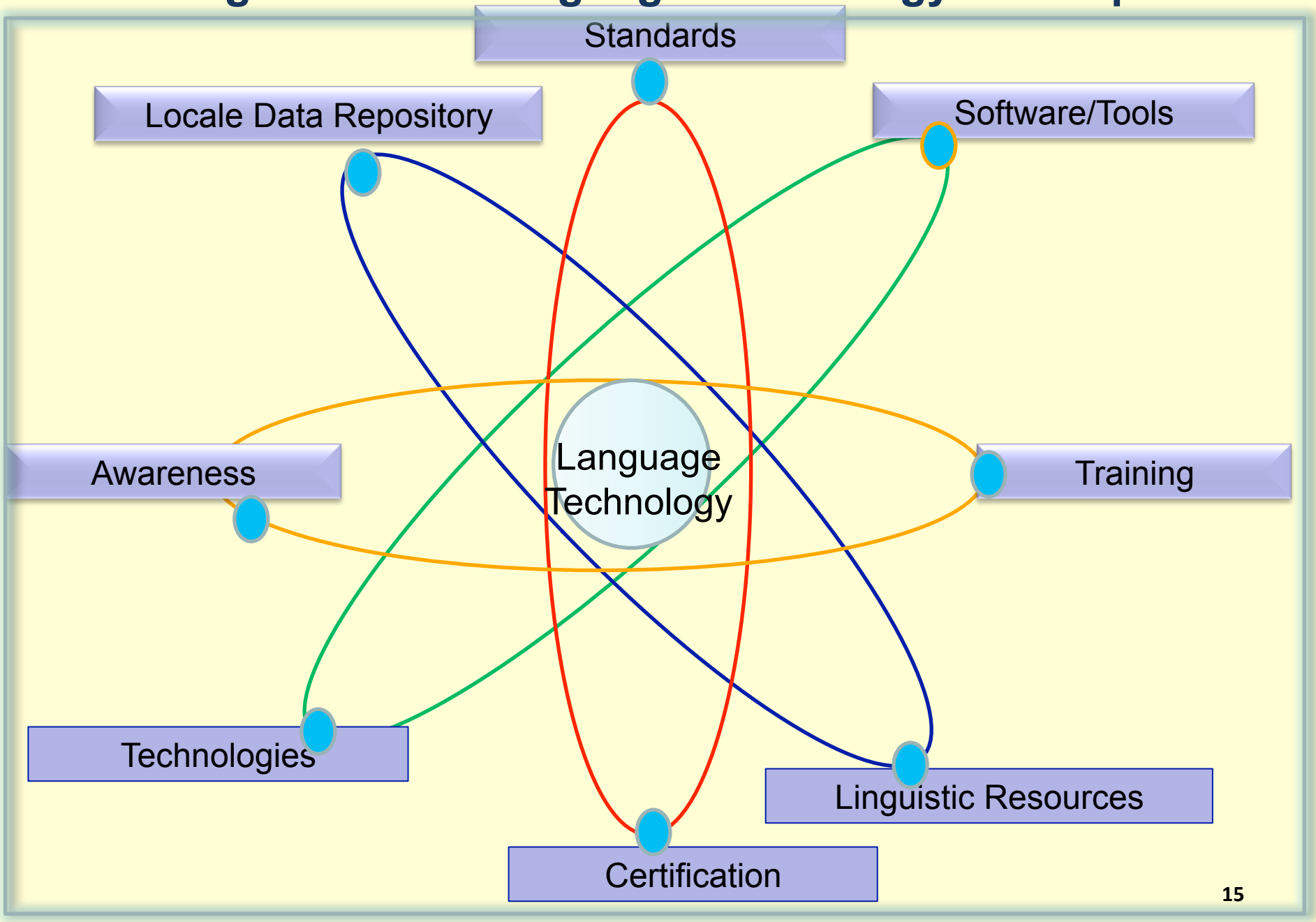
Local Language Interface – Not a desirable but An essential Component

- 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.

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

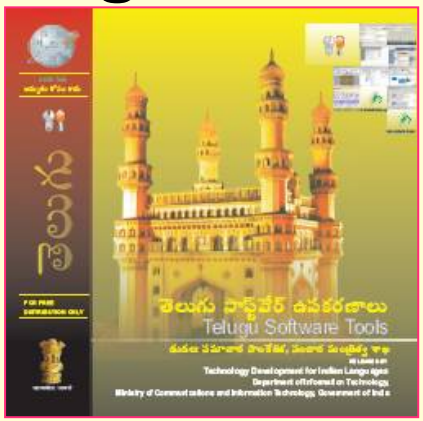
- Registered users of www.ildc.in web site of TDIL, DIT – through postal department.
- IT magazines, publications, etc.
- Schools, Government departments, etc.

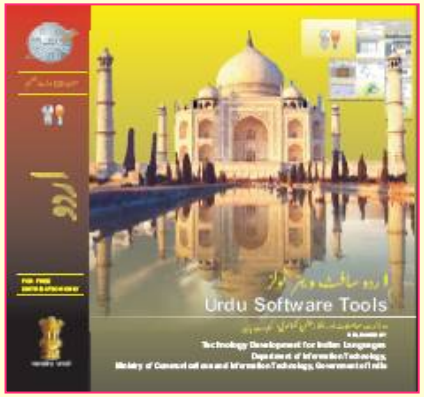
Software tools and fonts for 22 Indian languages namely Assamese, Bangla, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Manipuri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu languages have been released in public domain.

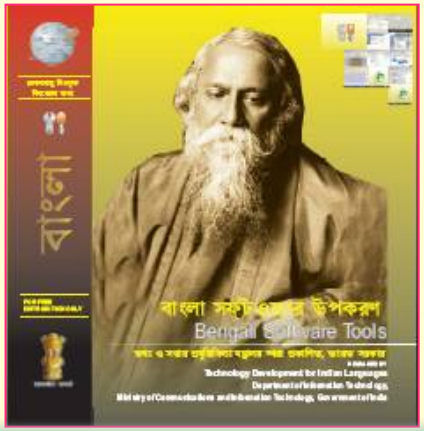
Freely downloadable from Indian Language Data centre – http://www.ildc.gov.in

Approx: 4.3 million downloads and 1.0 million shipments

CDs containing Indian Language Software Tools







Software tools and fonts CD contents

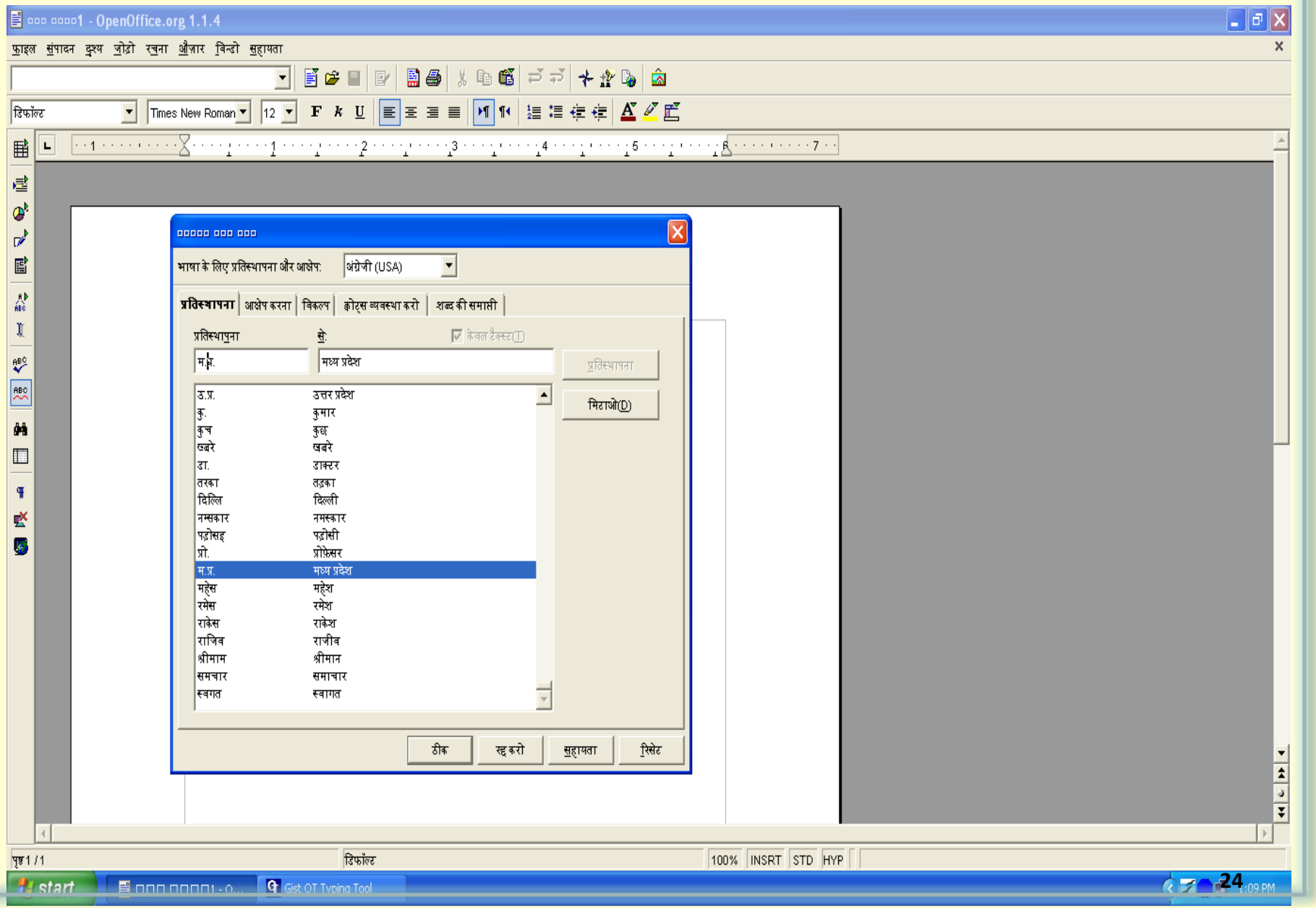
Common user –

- 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

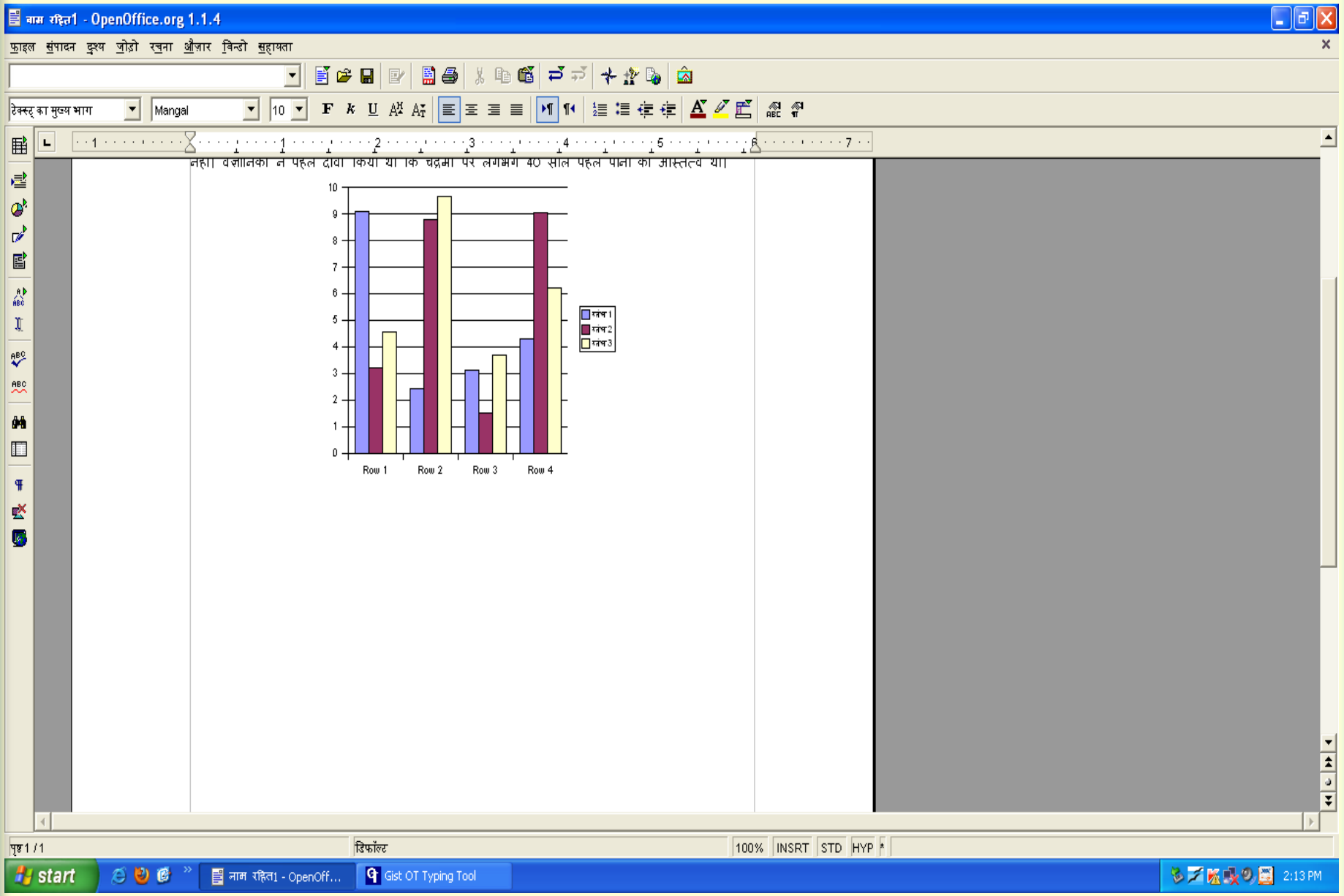
Power User –

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

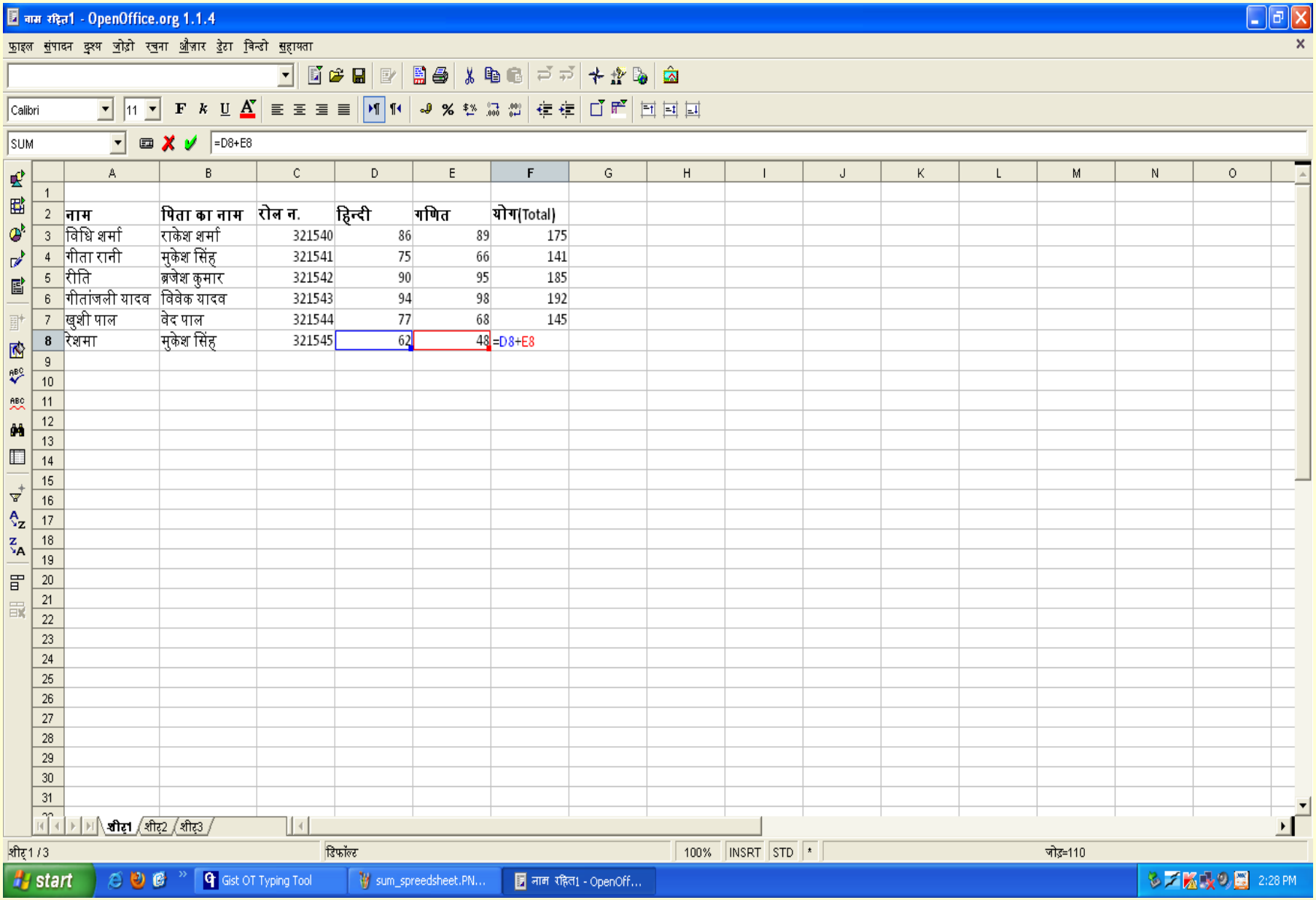
Screen shots of Localized Bharatiyaa Open Office - autocorrect



Screen shots of Localized Bharatiyaa Open Office - spreadsheet

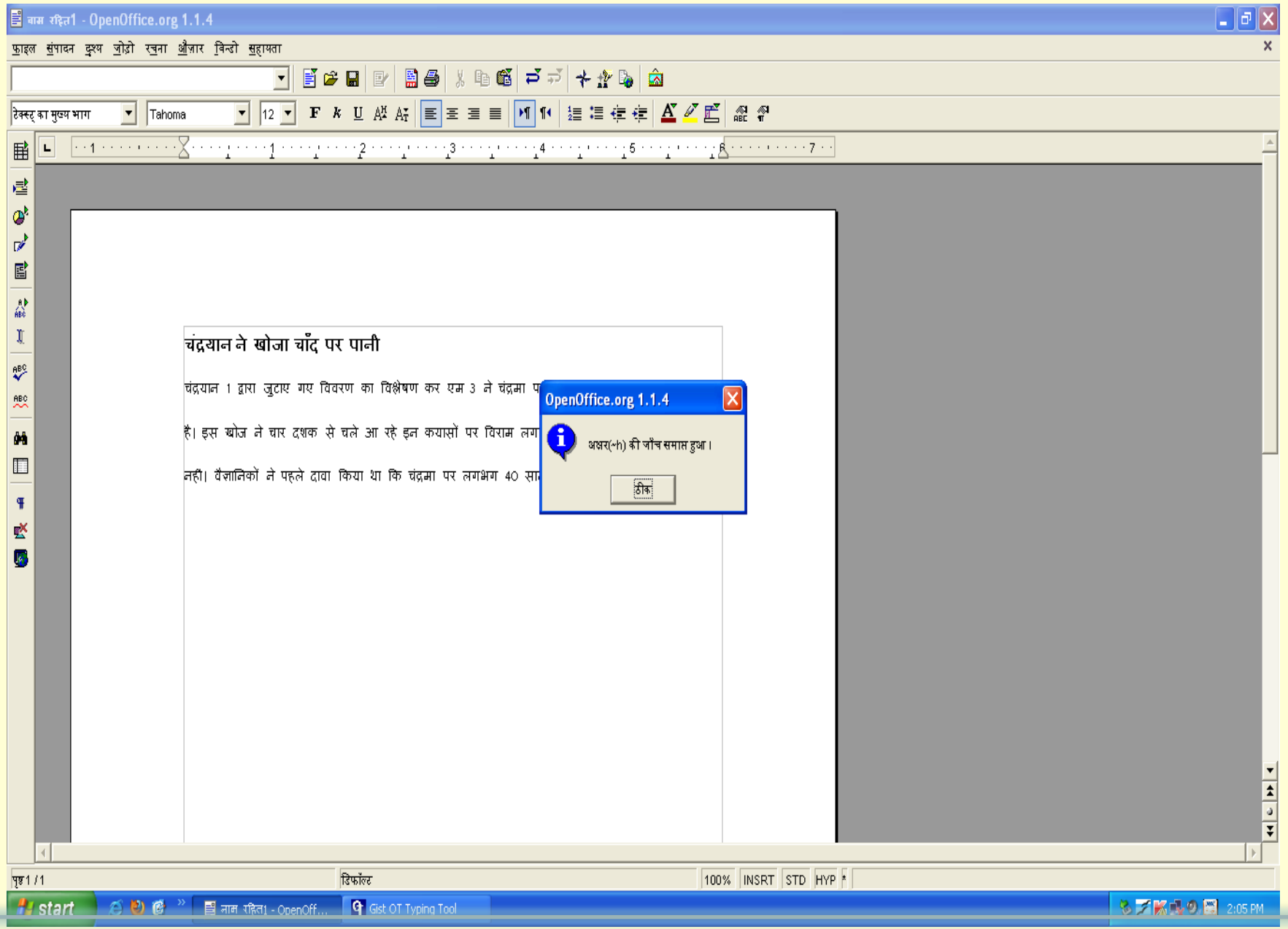


Screen shots of Localized Bharatiyaa Open Office - autosum

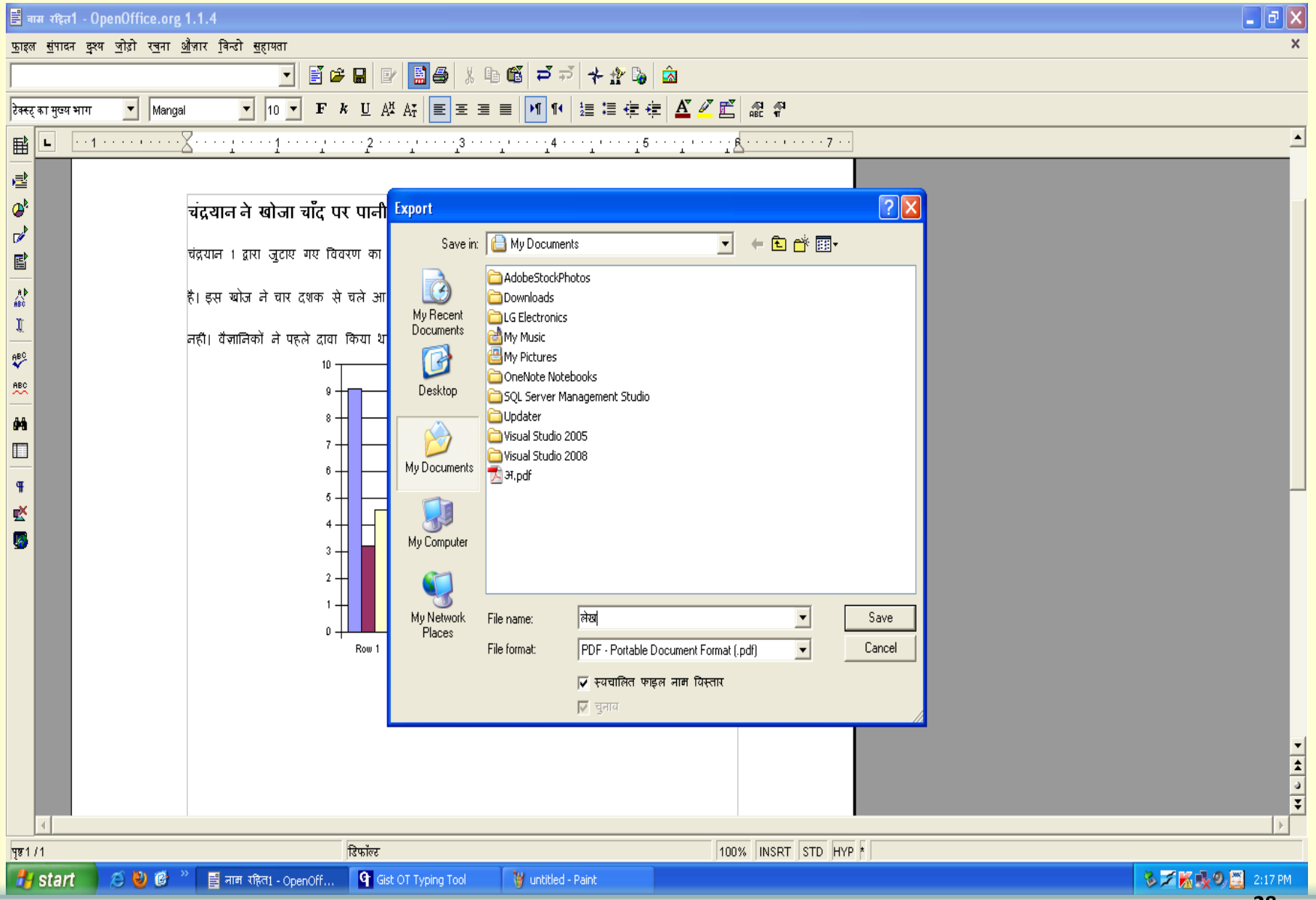


	A	B	C	D	E	F
1						
2	नाम	पिता का नाम	रोल न.	हिन्दी	गणित	योग(Total)
3	विधि शर्मा	राकेश शर्मा	321540	86	89	175
4	गीता रानी	मुकेश सिंह	321541	75	66	141
5	रीति	ब्रजेश कुमार	321542	90	95	185
6	गीतांजली यादव	विवेक यादव	321543	94	98	192
7	खुशी पाल	वेद पाल	321544	77	68	145
8	रेशमा	मुकेश सिंह	321545	62	48	=D8+E8

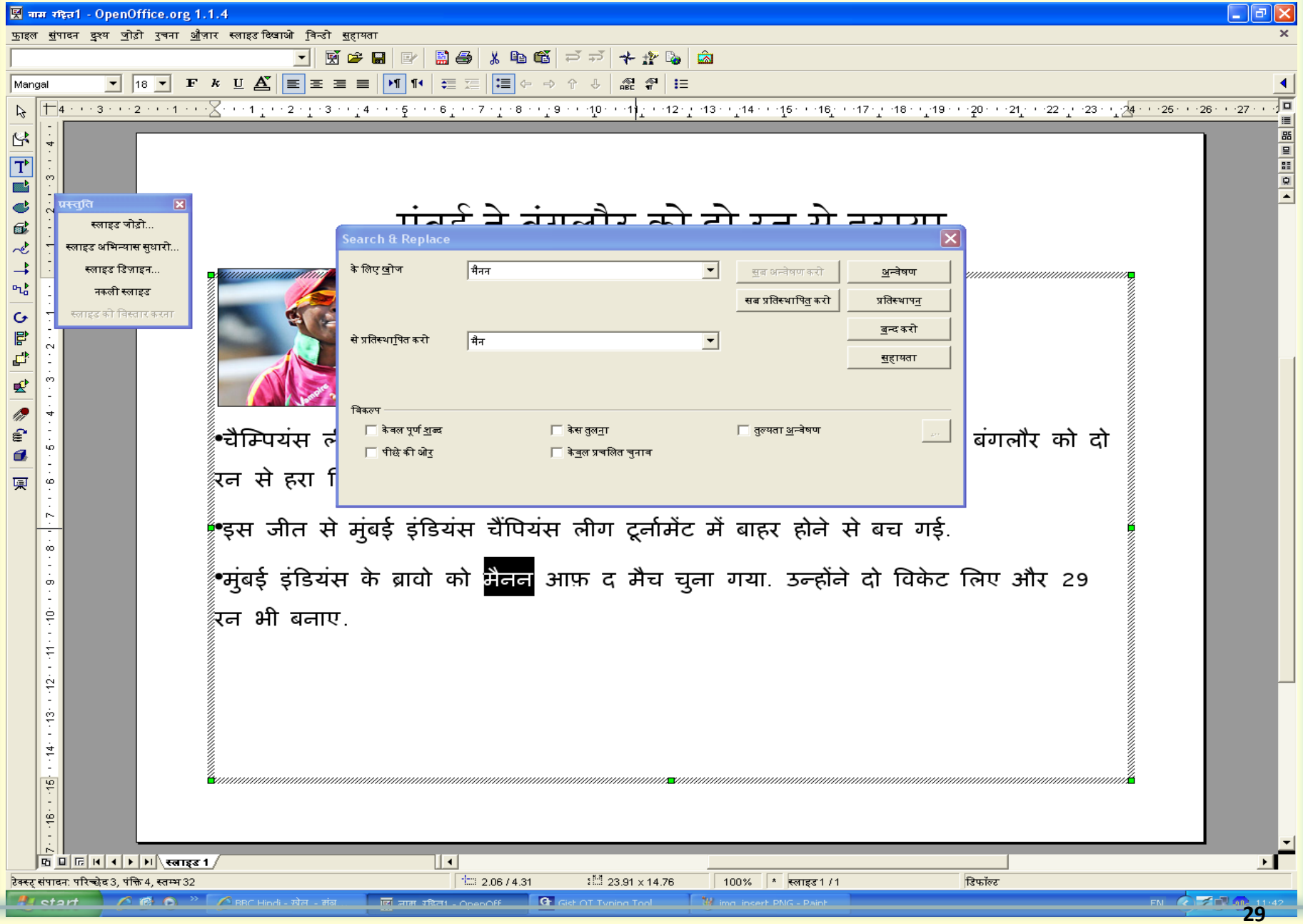
Screen shots of Localized Bharatiyaa Open Office (spell check tool)



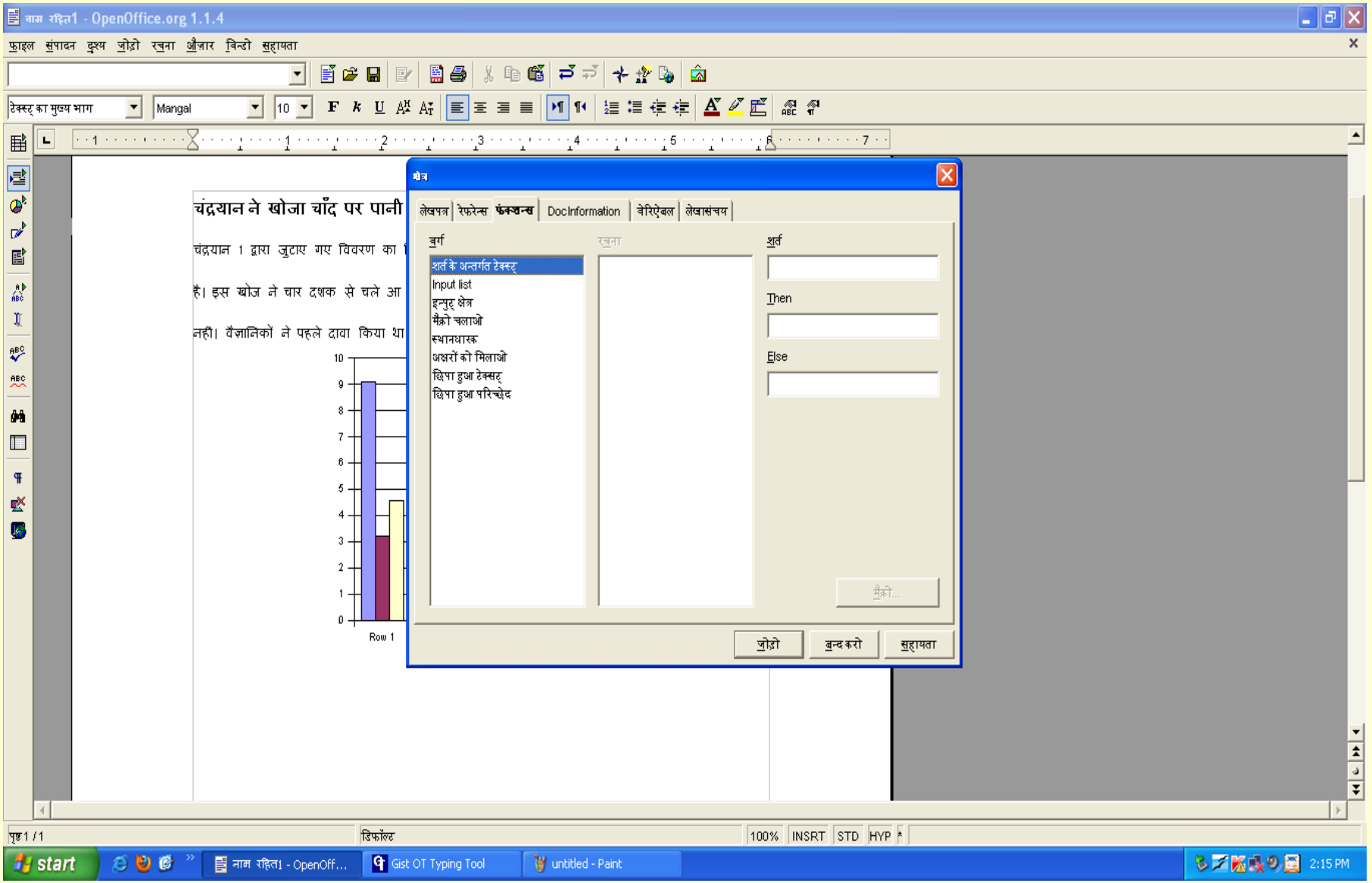
Screen shots of Localized Bharatiyaa Open Office – pdf converter



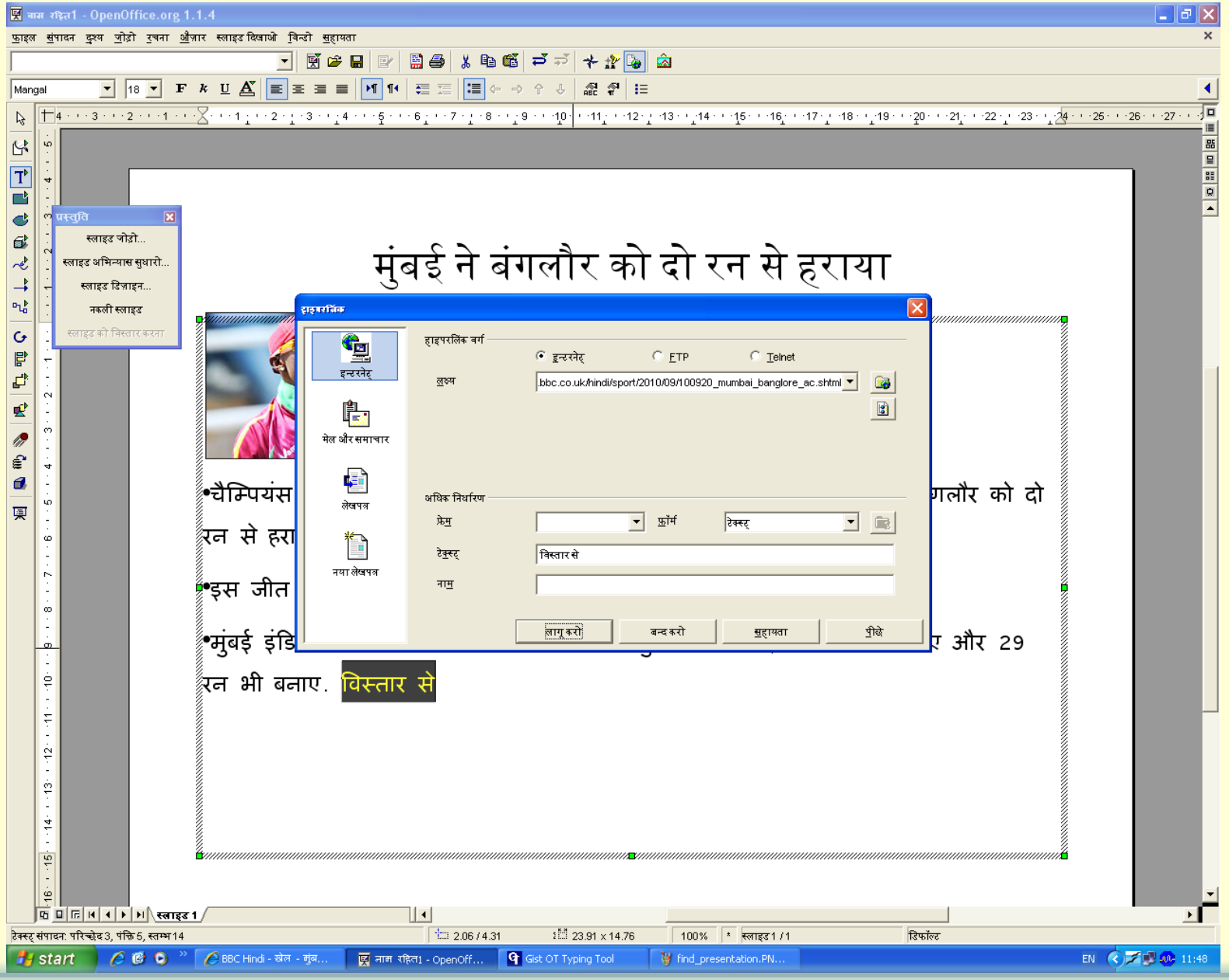
Screen shots of Localized Bharatiyaa Open Office – find & Replace



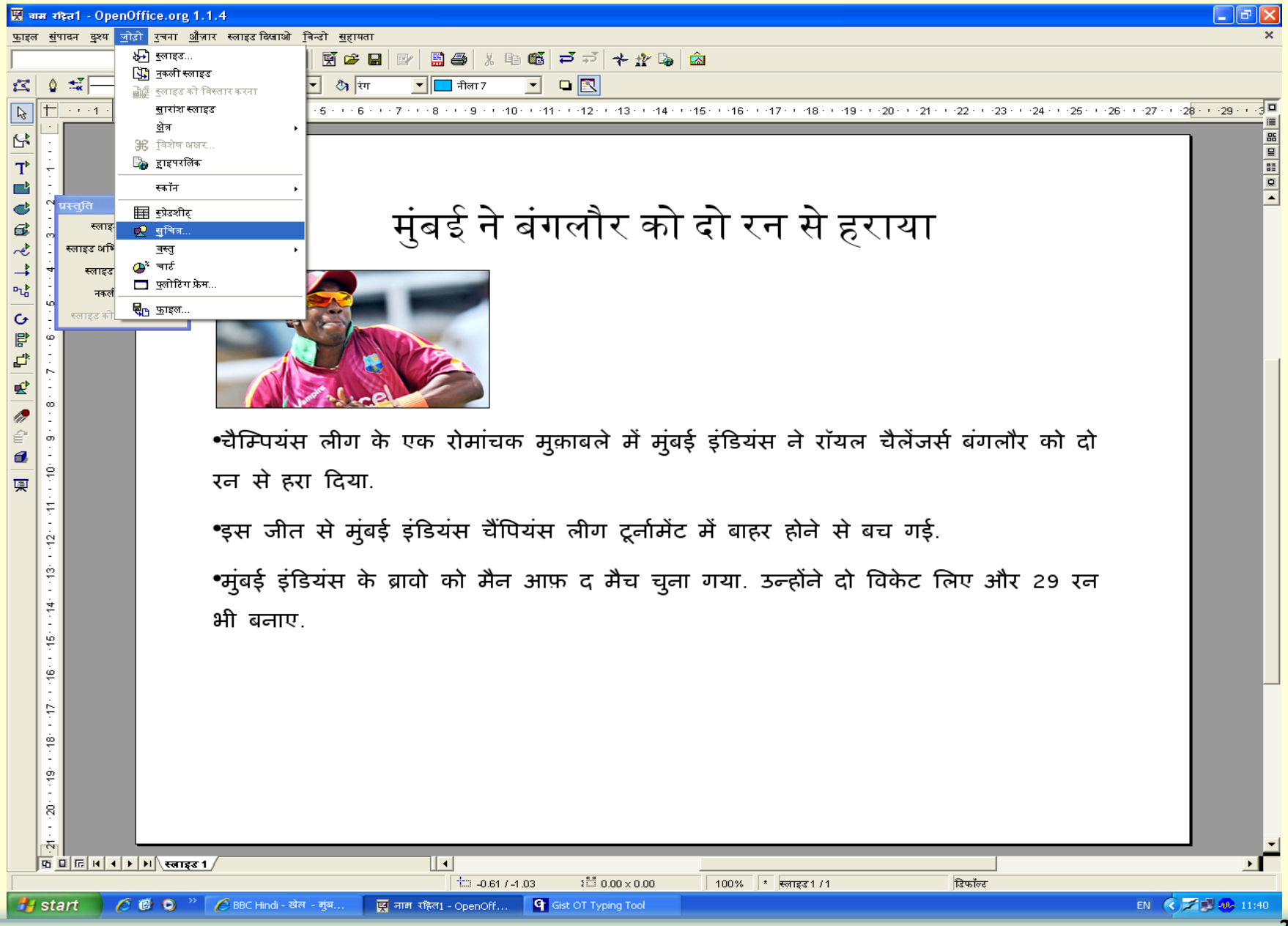
Screen shots of Localized Bharatiyaa Open Office



Screen shots of Localized Bharatiyaa Open Office –insert link



Screen shots of Localized Bharatiyaa Open Office



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.]

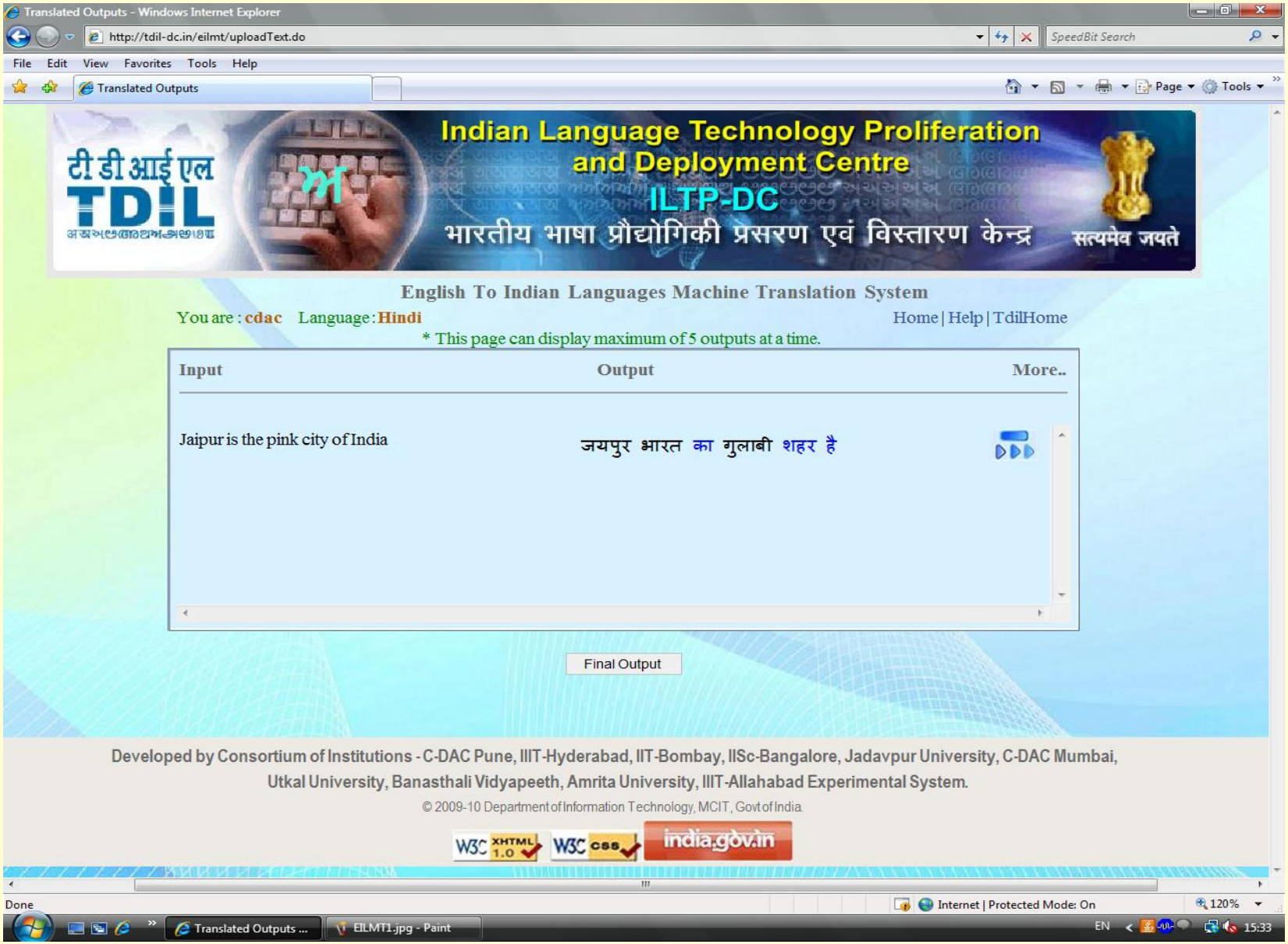
- Indian Languages to Indian Languages Machine Translation System
  [9 Language Pairs: Telugu-Hindi, Hindi-Tamil, Urdu-Hindi, Kannada-Hindi, Punjabi-Hindi, Marathi-Hindi, Bengali-Hindi, Tamil-Telugu, Malayalam-Tamil]

- Cross-Lingual Information Access
  [6 Languages : Hindi , Bengali, Tamil , Marathi , Telugu and Punjabi]

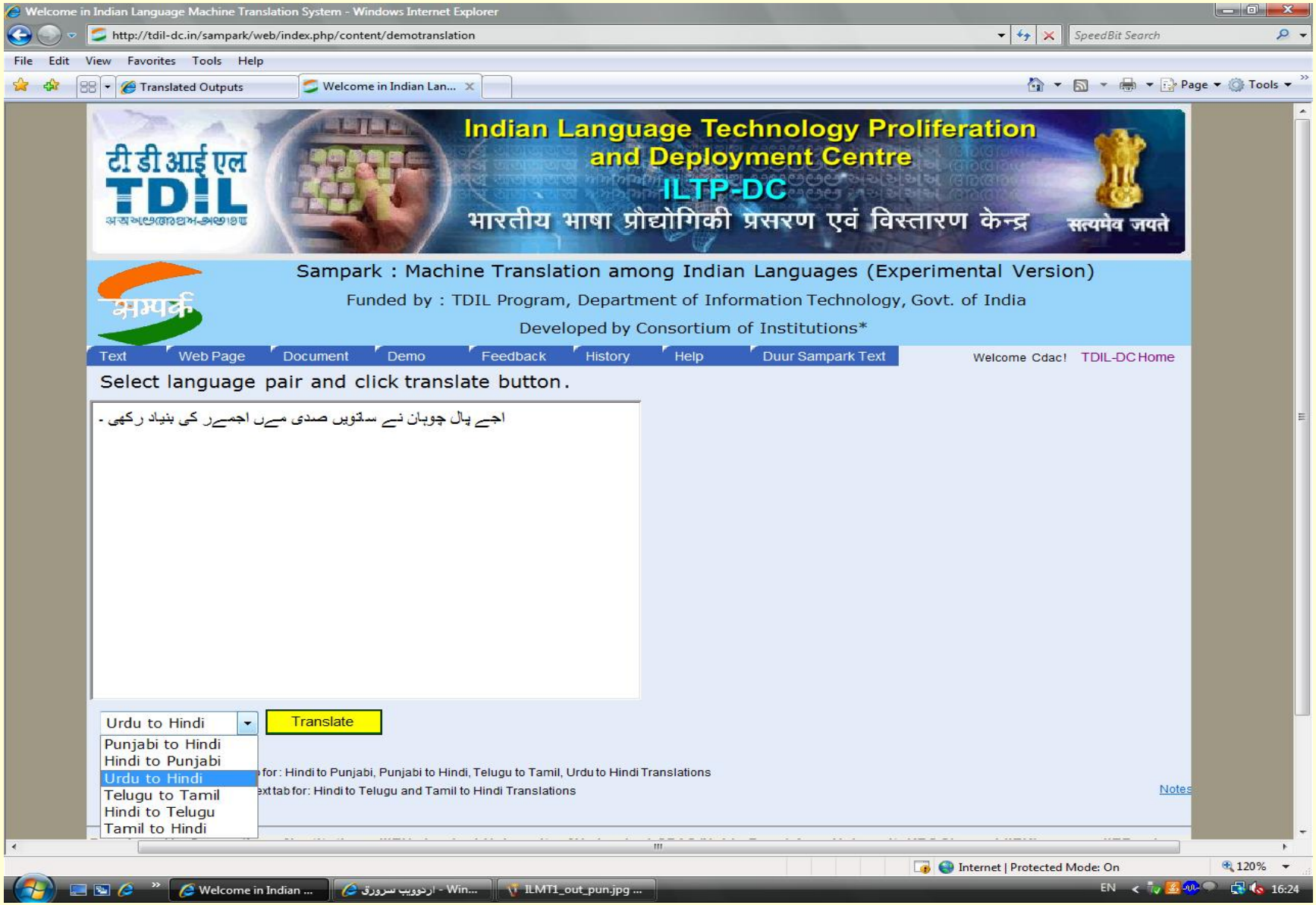
- Optical Character Recognition Systems
  [10 Scripts: Bangla, Devnagari, Malayalam, Gujrati, Tamil, Telugu, Kannada, Oriya, Gurumukhi, Tibetan]

- On-line Handwriting recognition system [6 Scripts: Hindi , Bengali , Tamil , Telugu , Kannada and Malayalam]

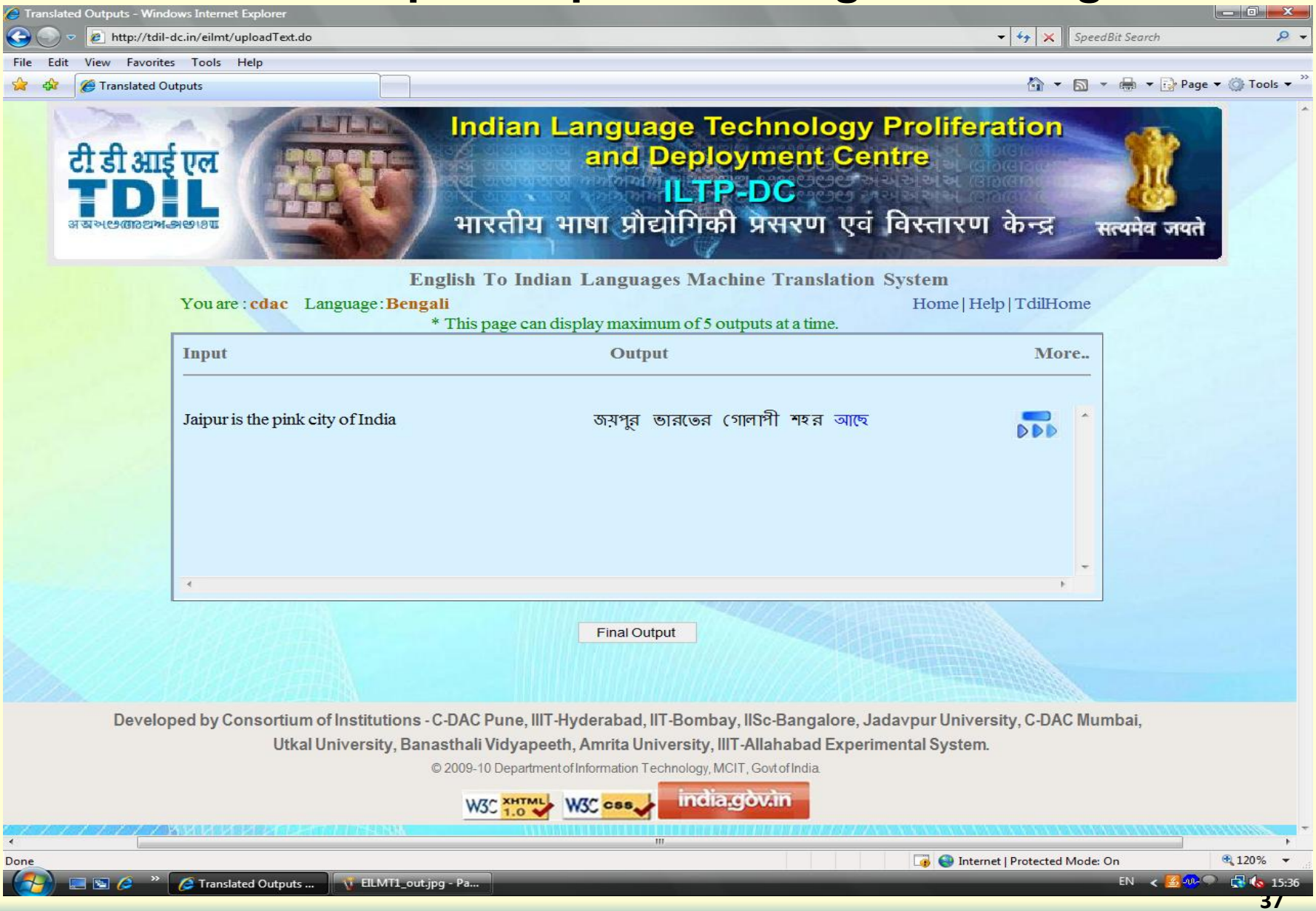
Sample Outputs For English - Hindi



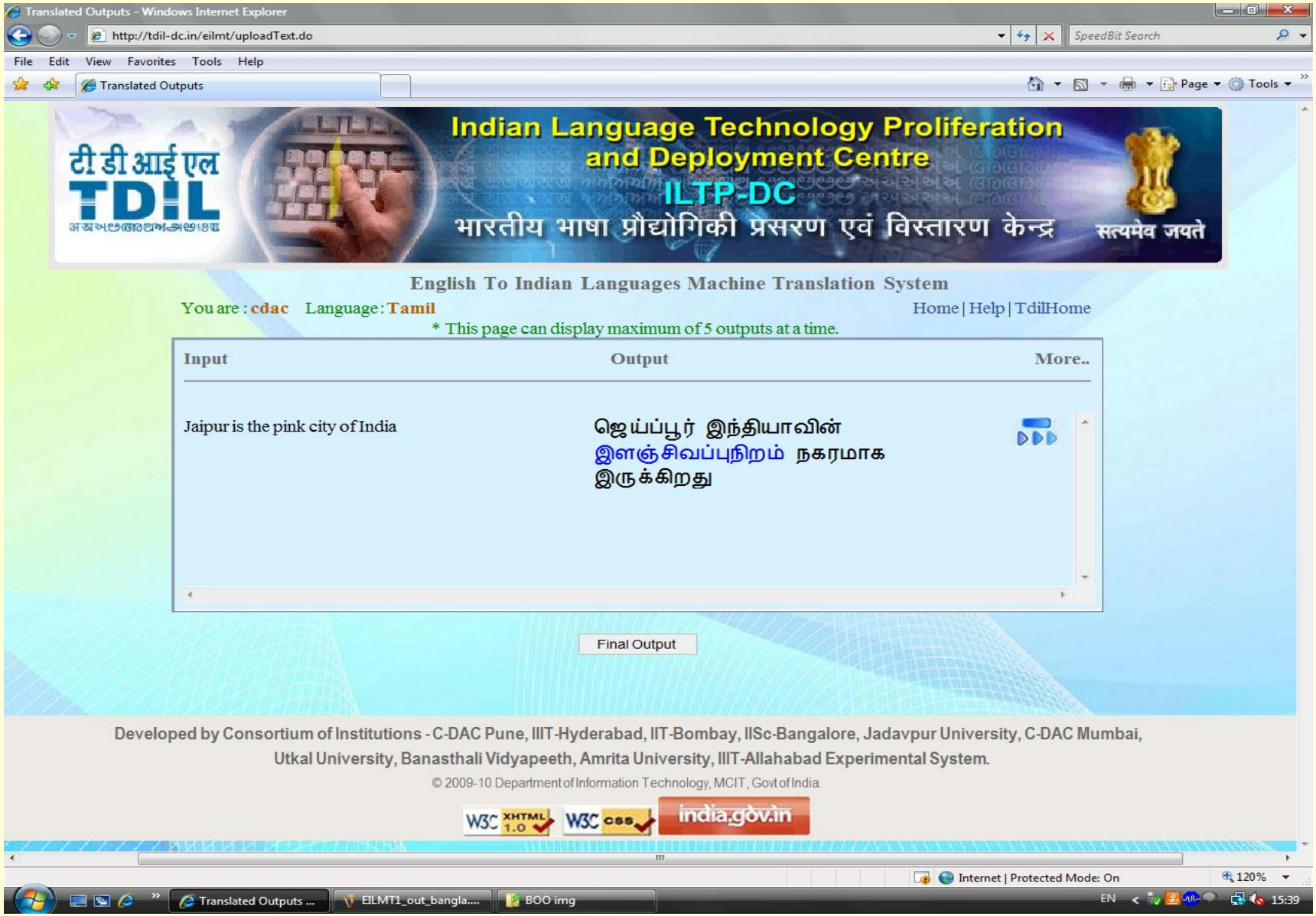
Sample Outputs For English - Urdu



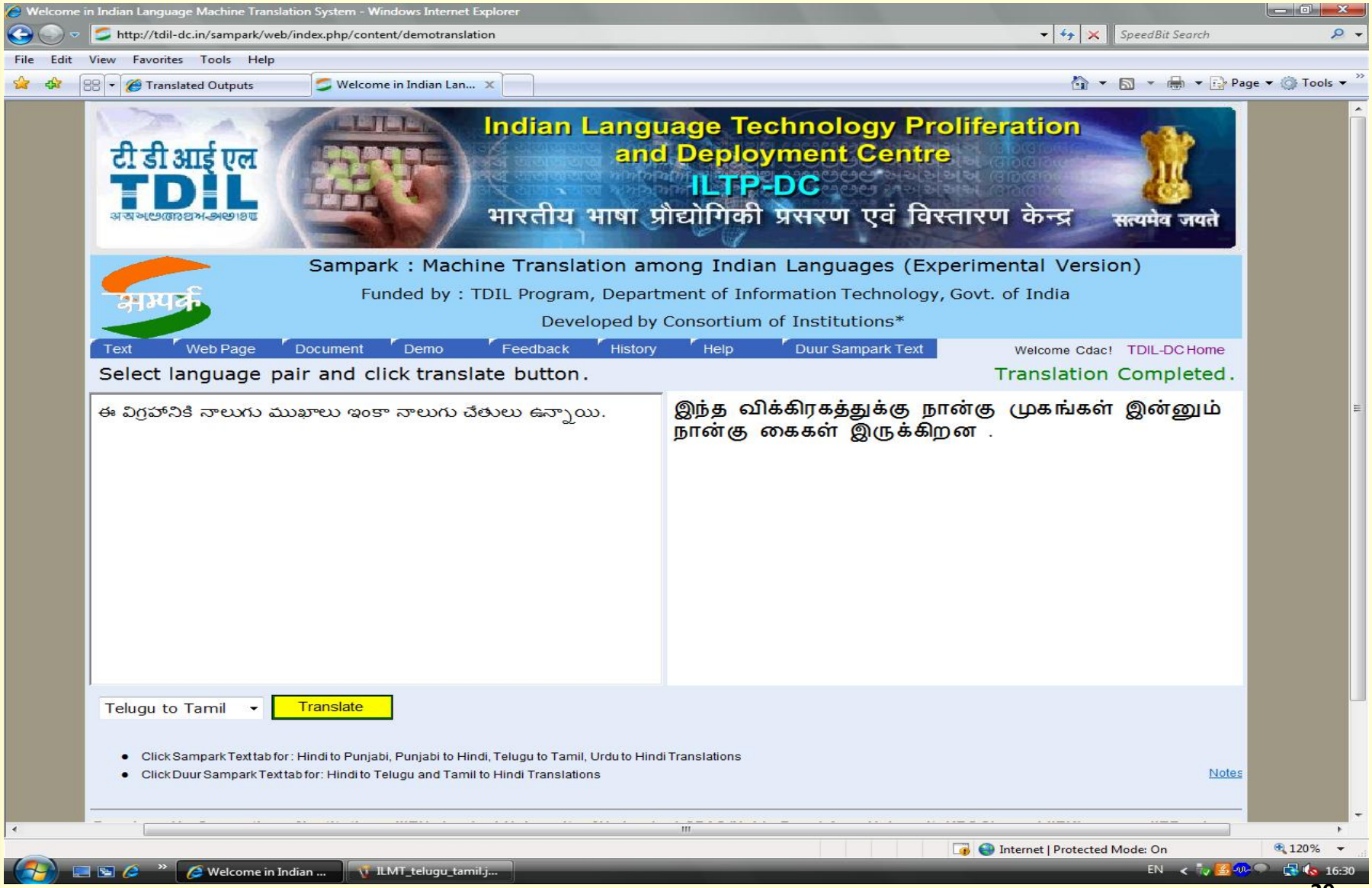
Sample Outputs For English - Bangla



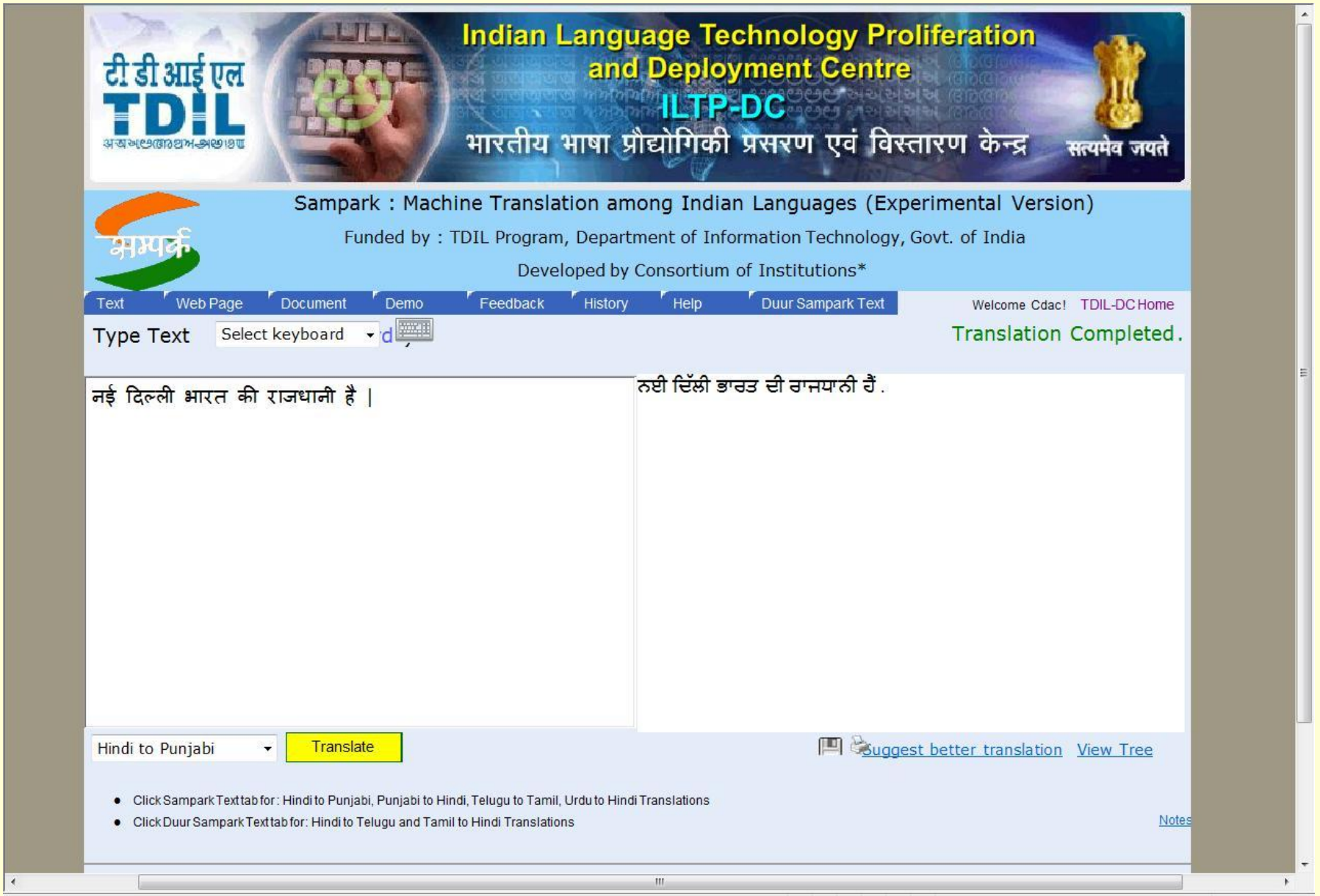
Sample Outputs For English - Tamil



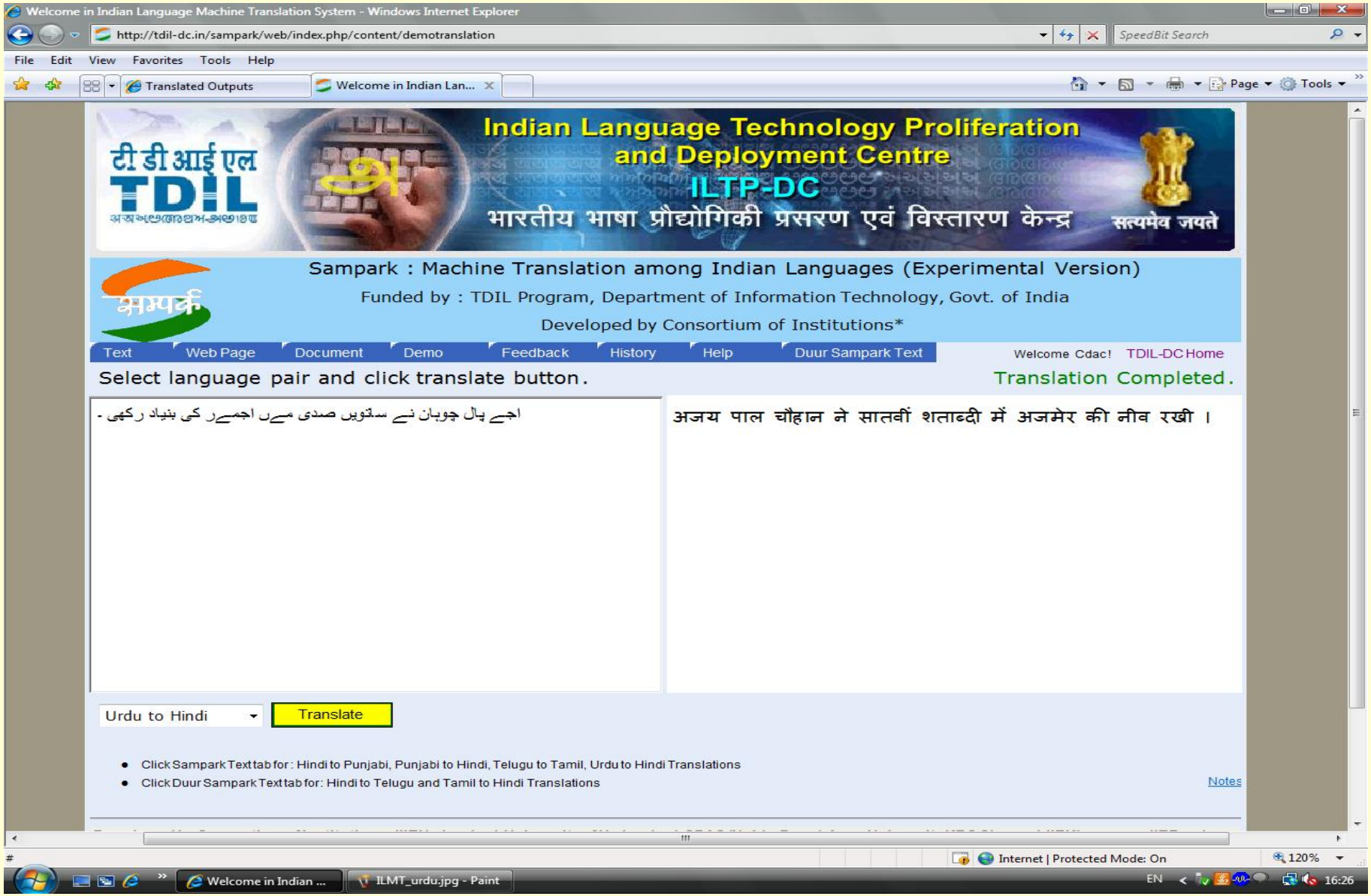
Indian Language to Indian Languages Machine Translation System



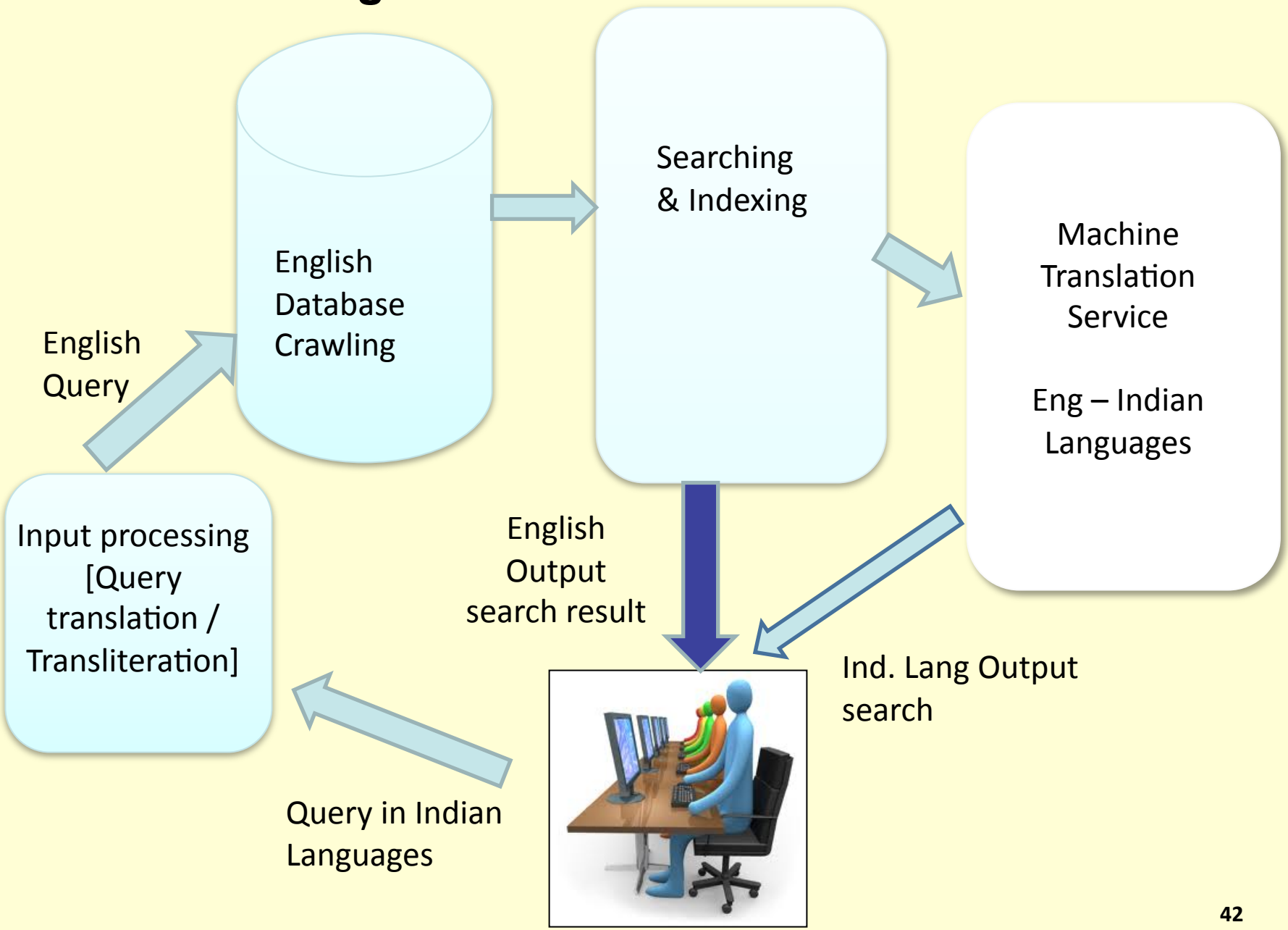
Sample Outputs For Hindi -Punjabi



Urdu-Hindi ILMT Translation through URL :- Urdu input

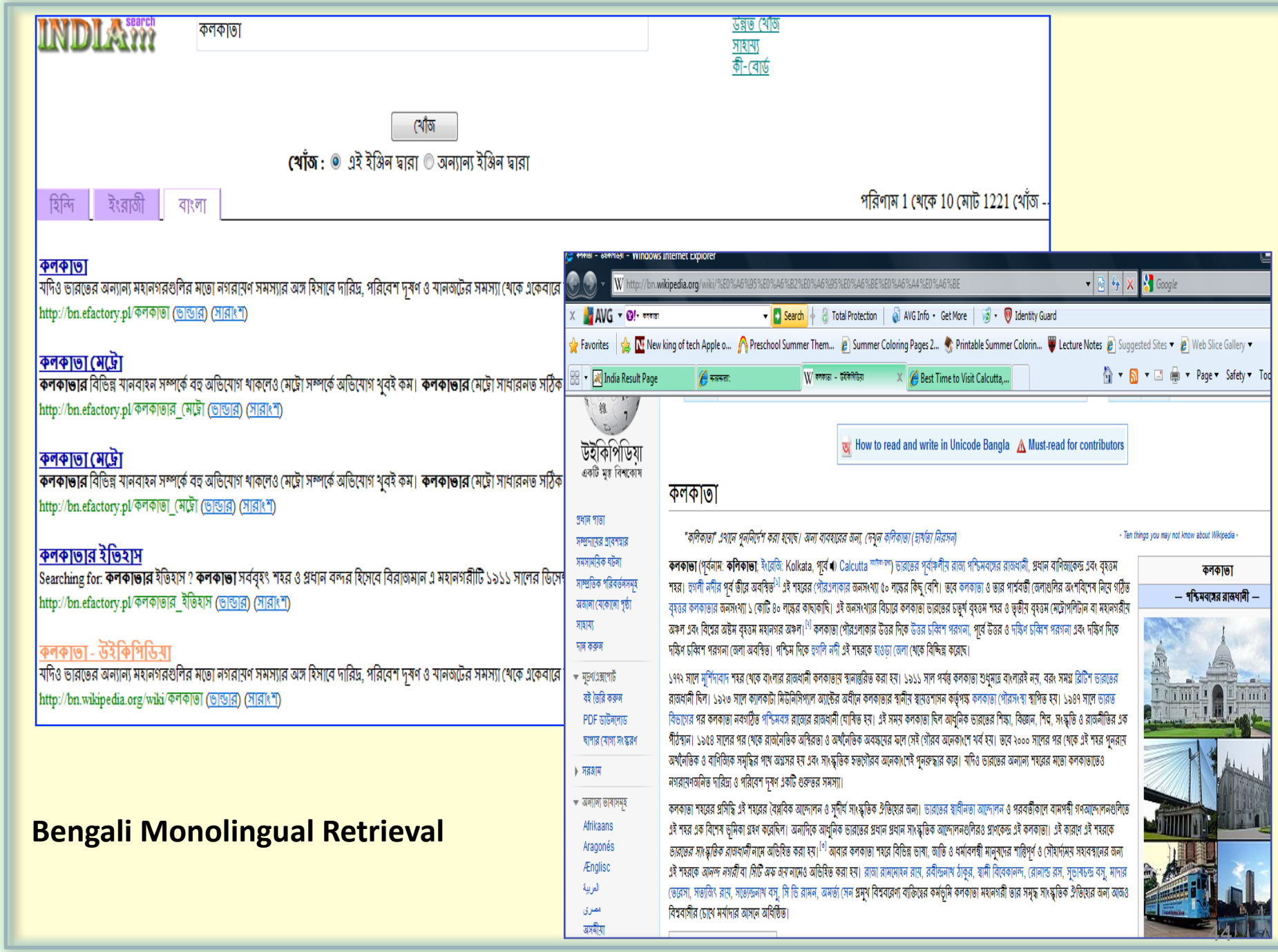


CLIA integrated with Machine Translation

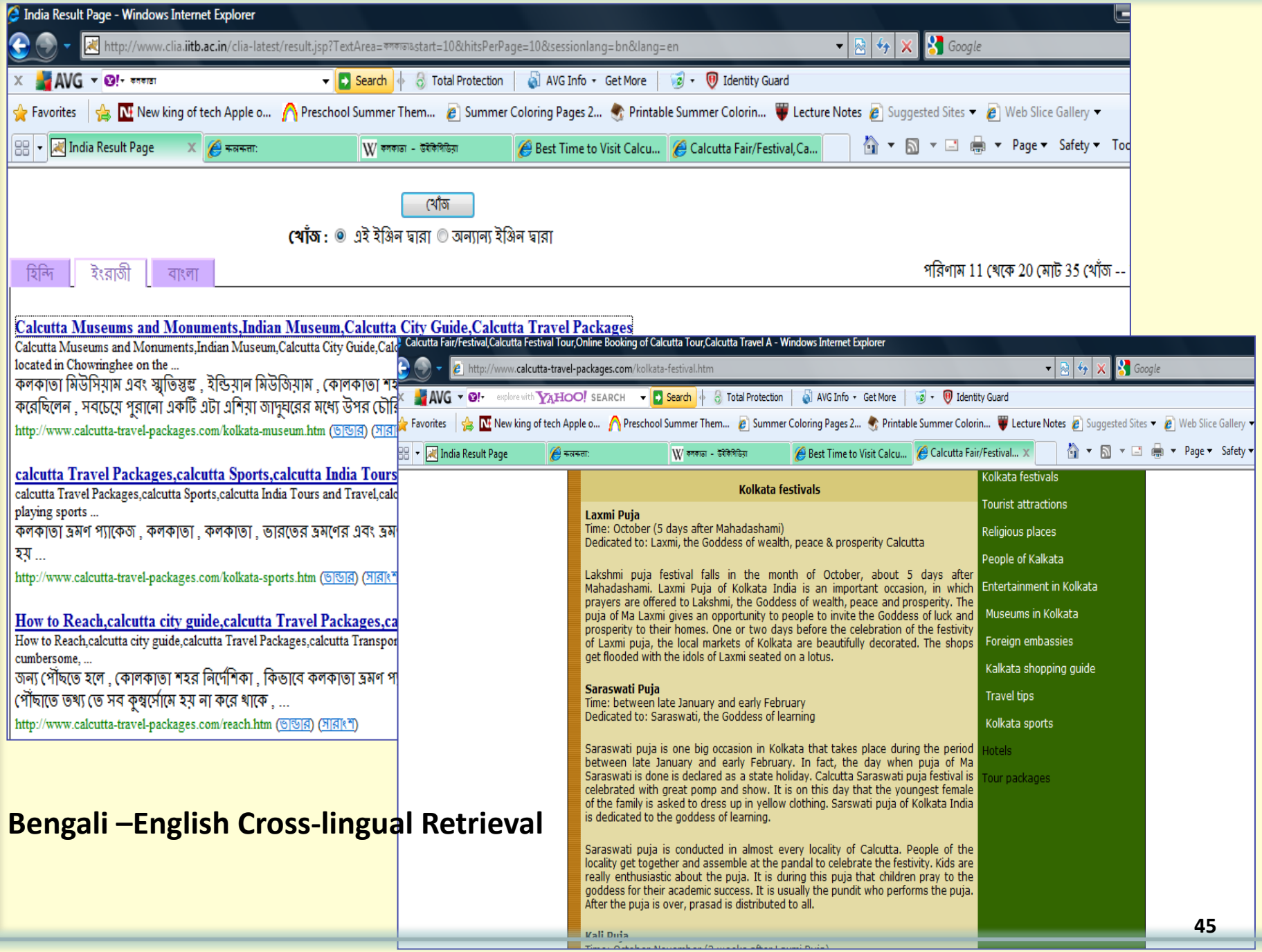


Cross-Lingual Information Access (CLIA)

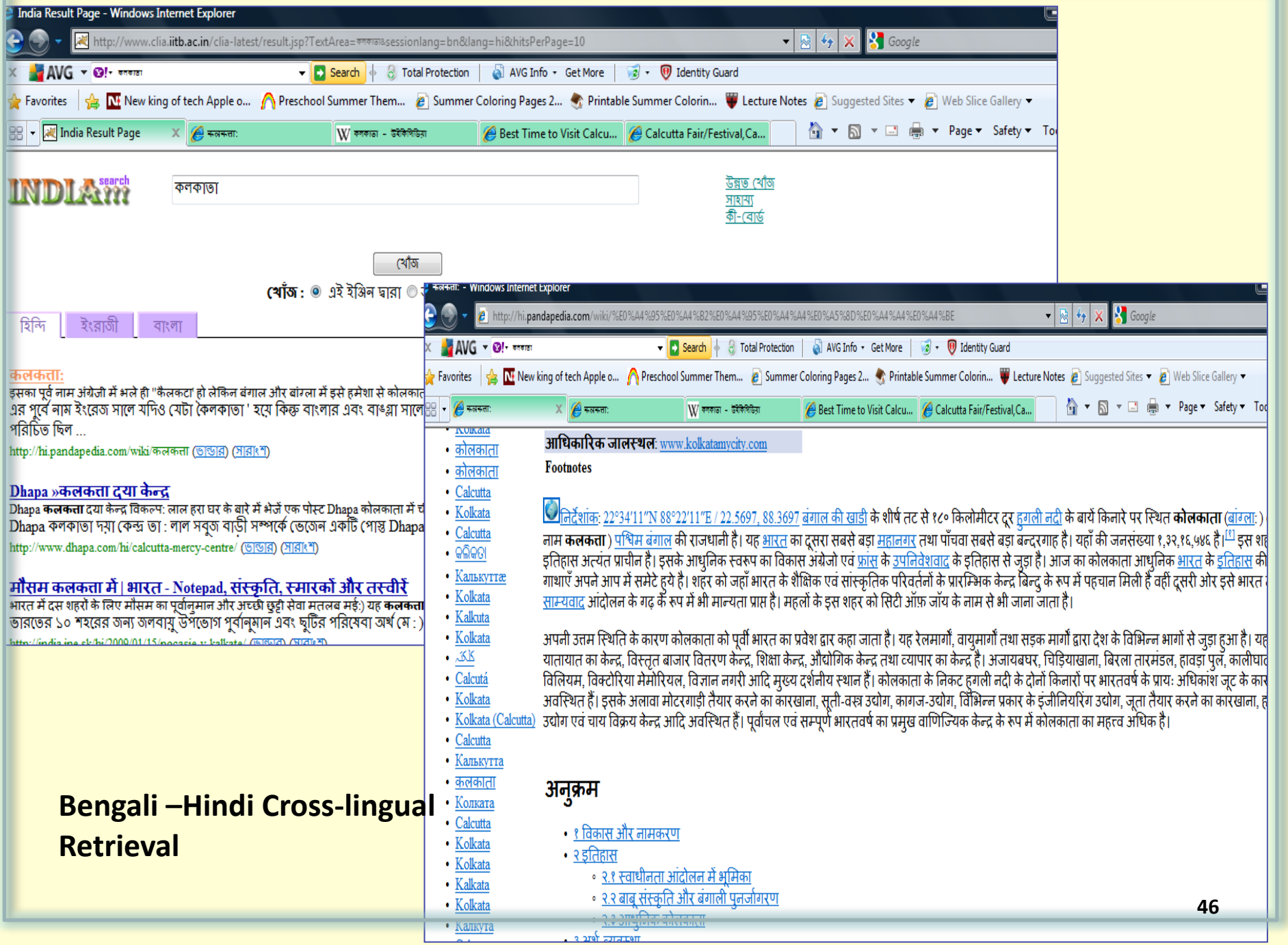
- In CLIA, the input query is in one language and information is retrieved in another.
- The query language is one of Bangla, Hindi, Marathi, Punjabi, Tamil and Telugu.
- The retrieved documents are in English, Hindi or the language of the query.



Bengali Monolingual Retrieval

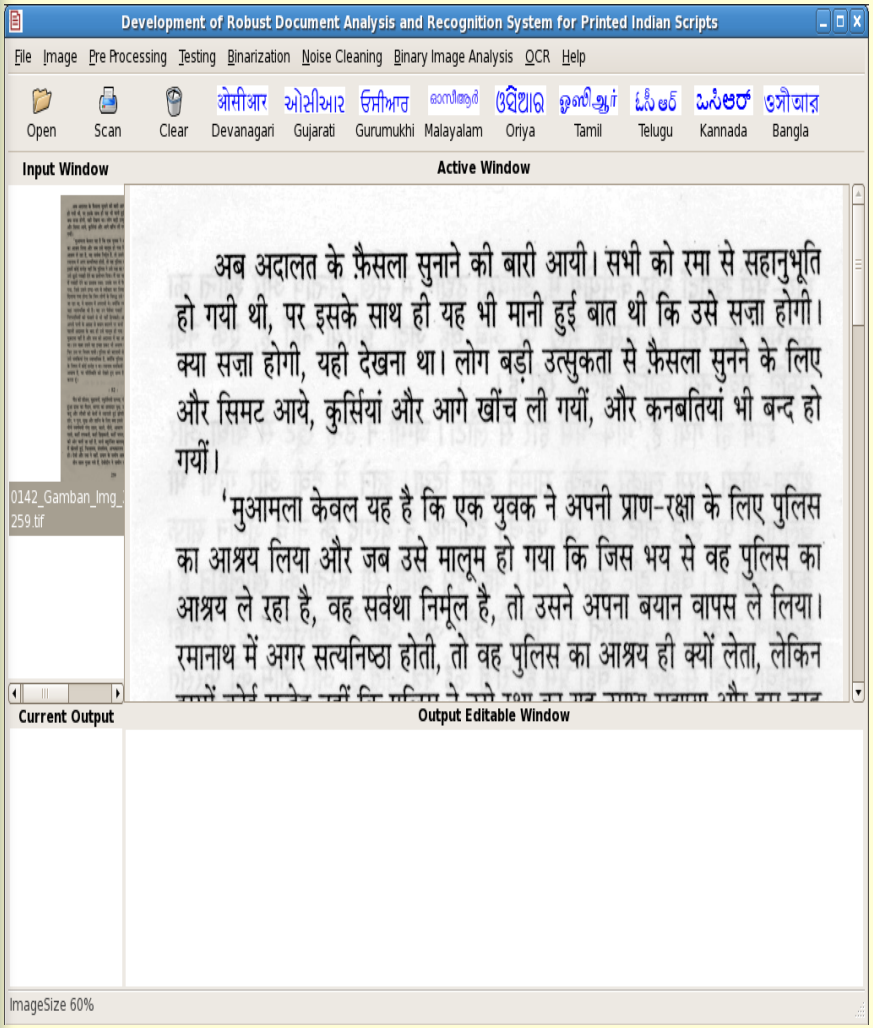


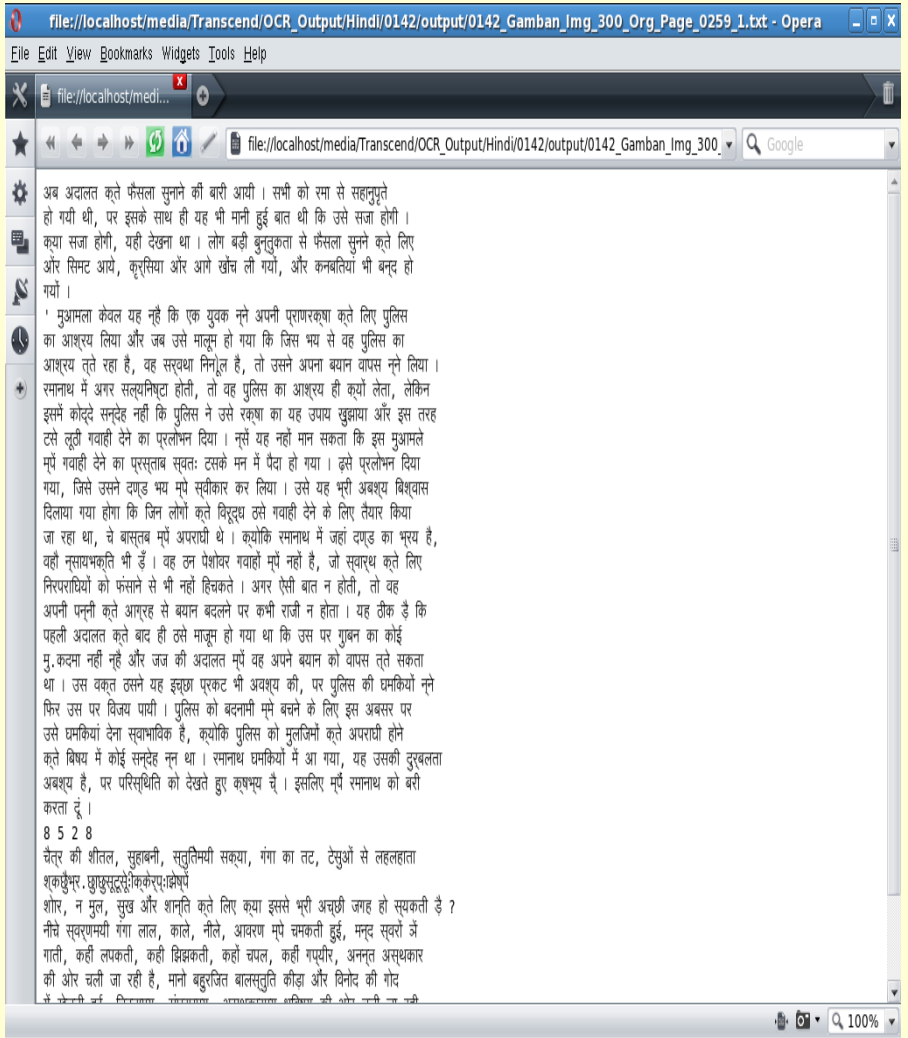
Bengali –English Cross-lingual Retrieval



Bengali –Hindi Cross-lingual Retrieval

OCR

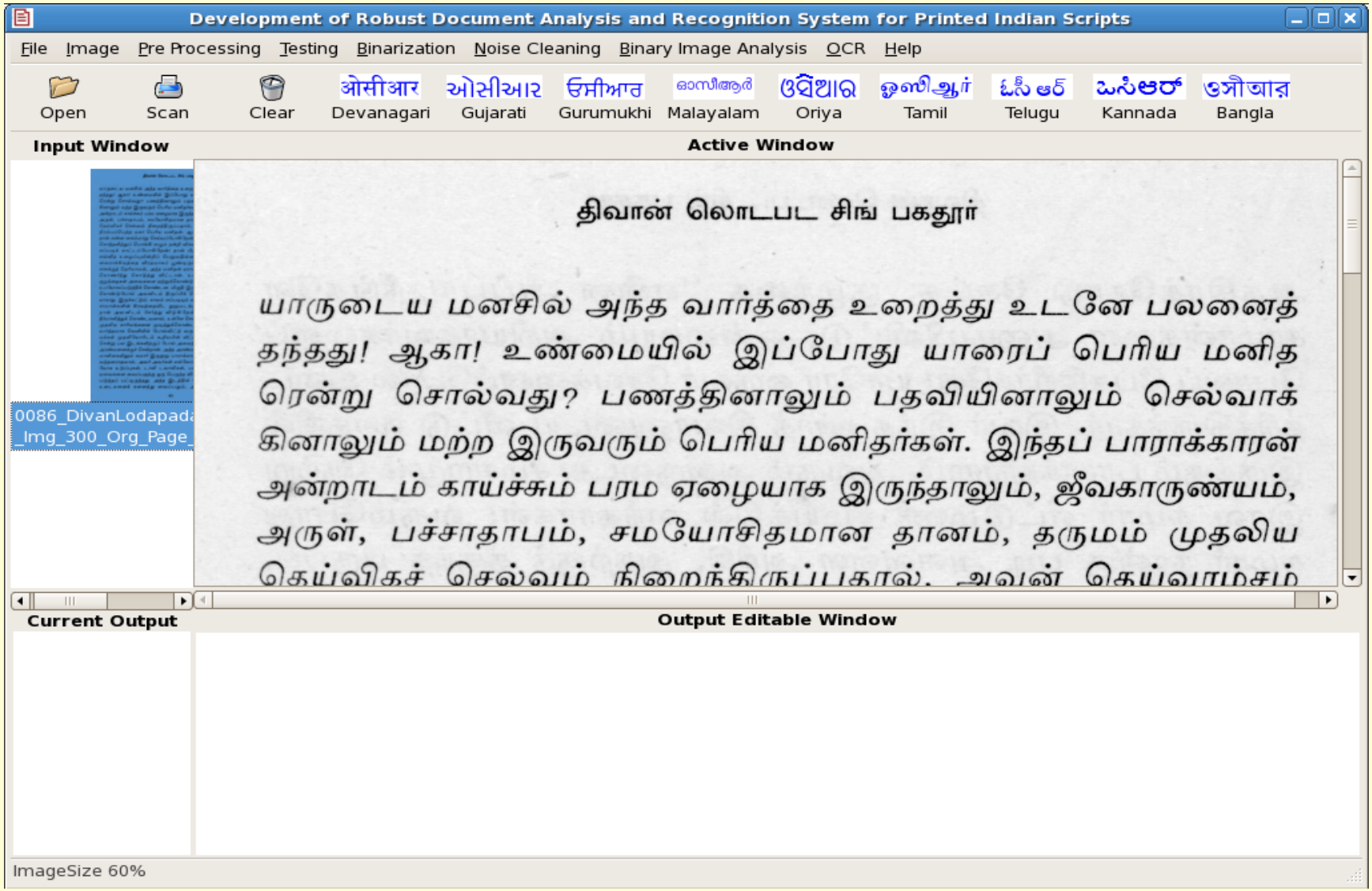




Original Image Loaded in Input Window of OCR

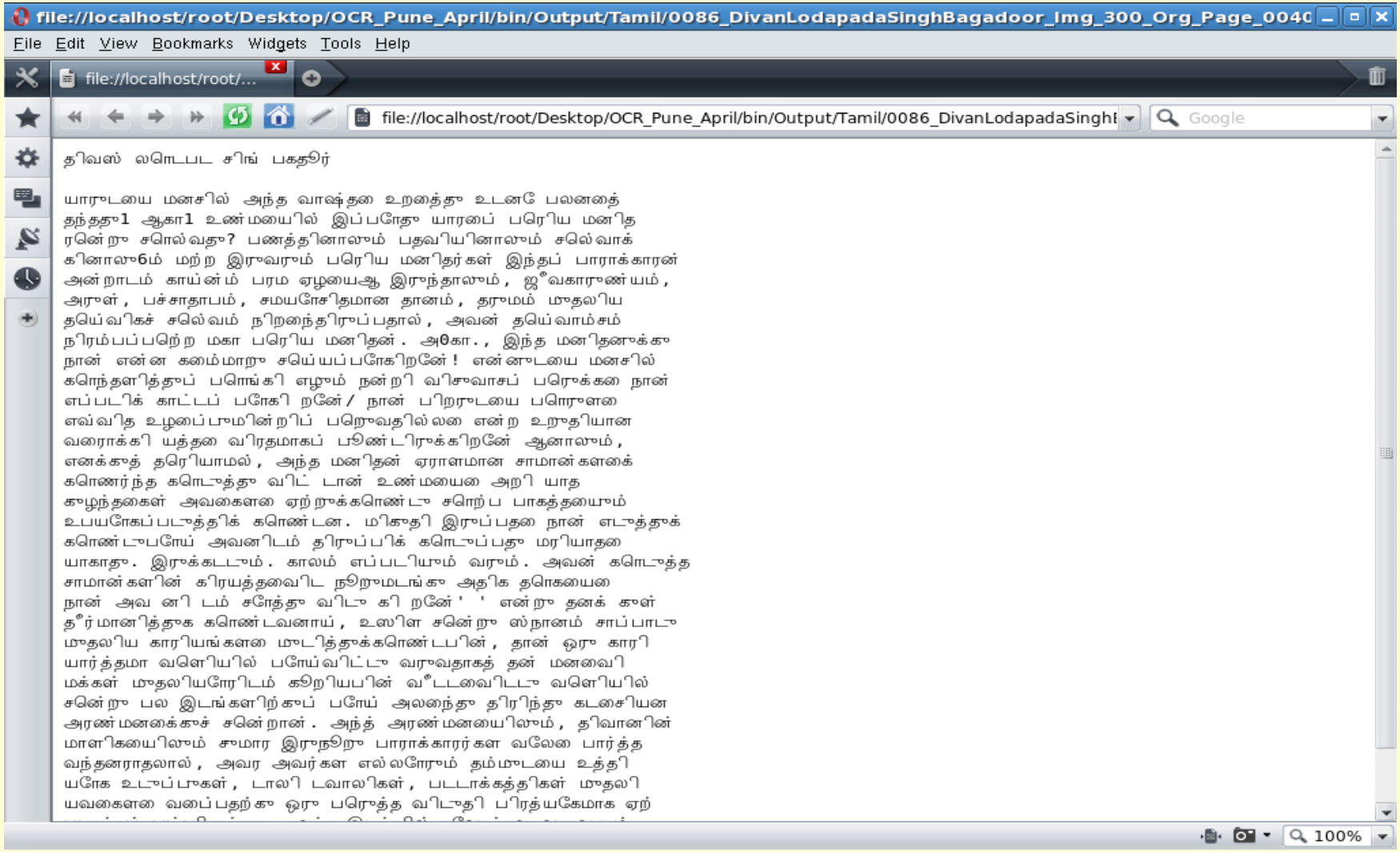
Output of Devnagari OCR

OCR



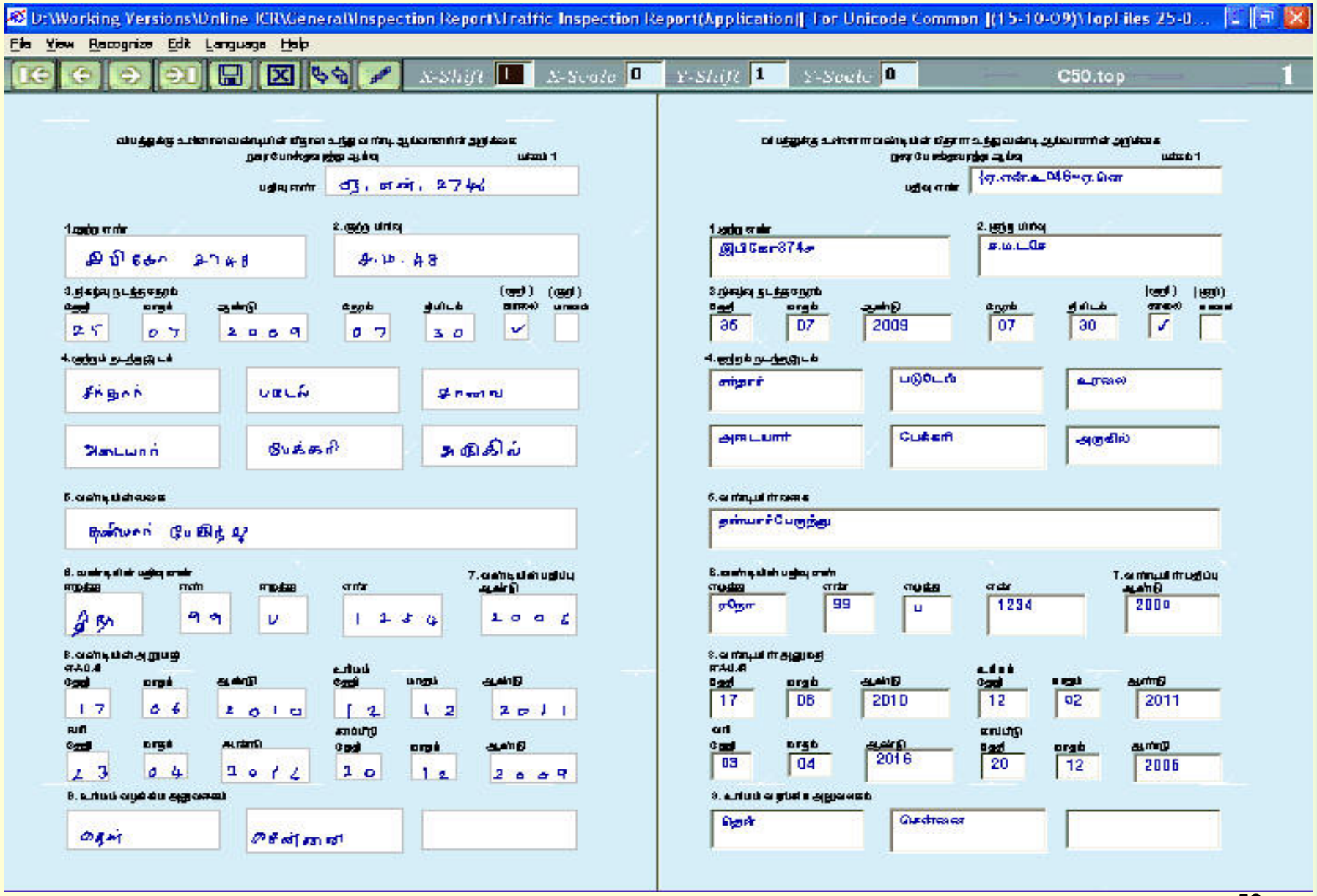
Original Image Loaded in Input Window of OCR

OCR



Output of Tamil OCR

Sample Tamil OHWR form



Technology development for inclusive growth

Text to Speech in Indian Languages

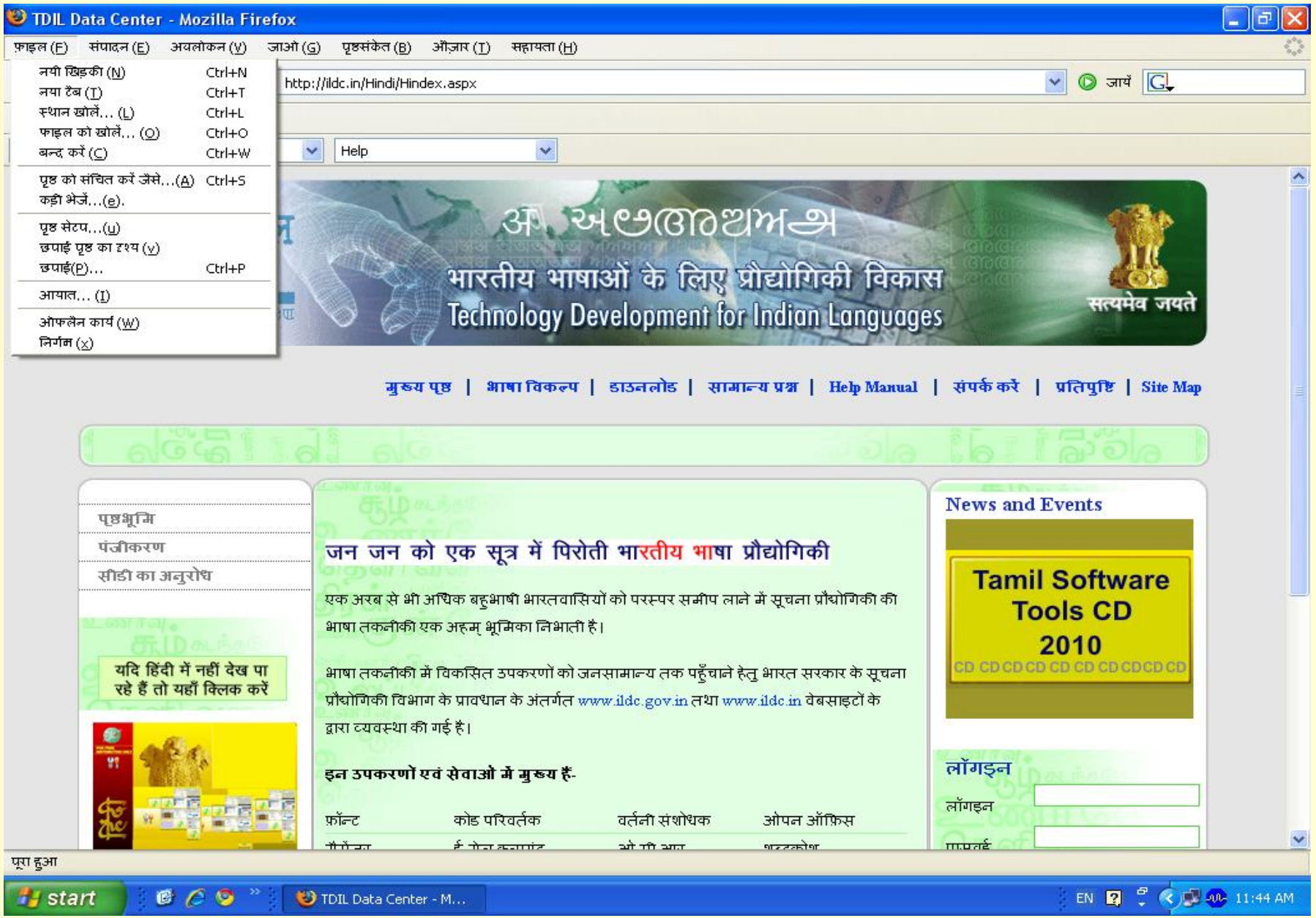
To develop accessible technologies for differently abled section of the society , TDIL **programme** has undertaken initiatives.

Consortium mode project has been initiated for development of:

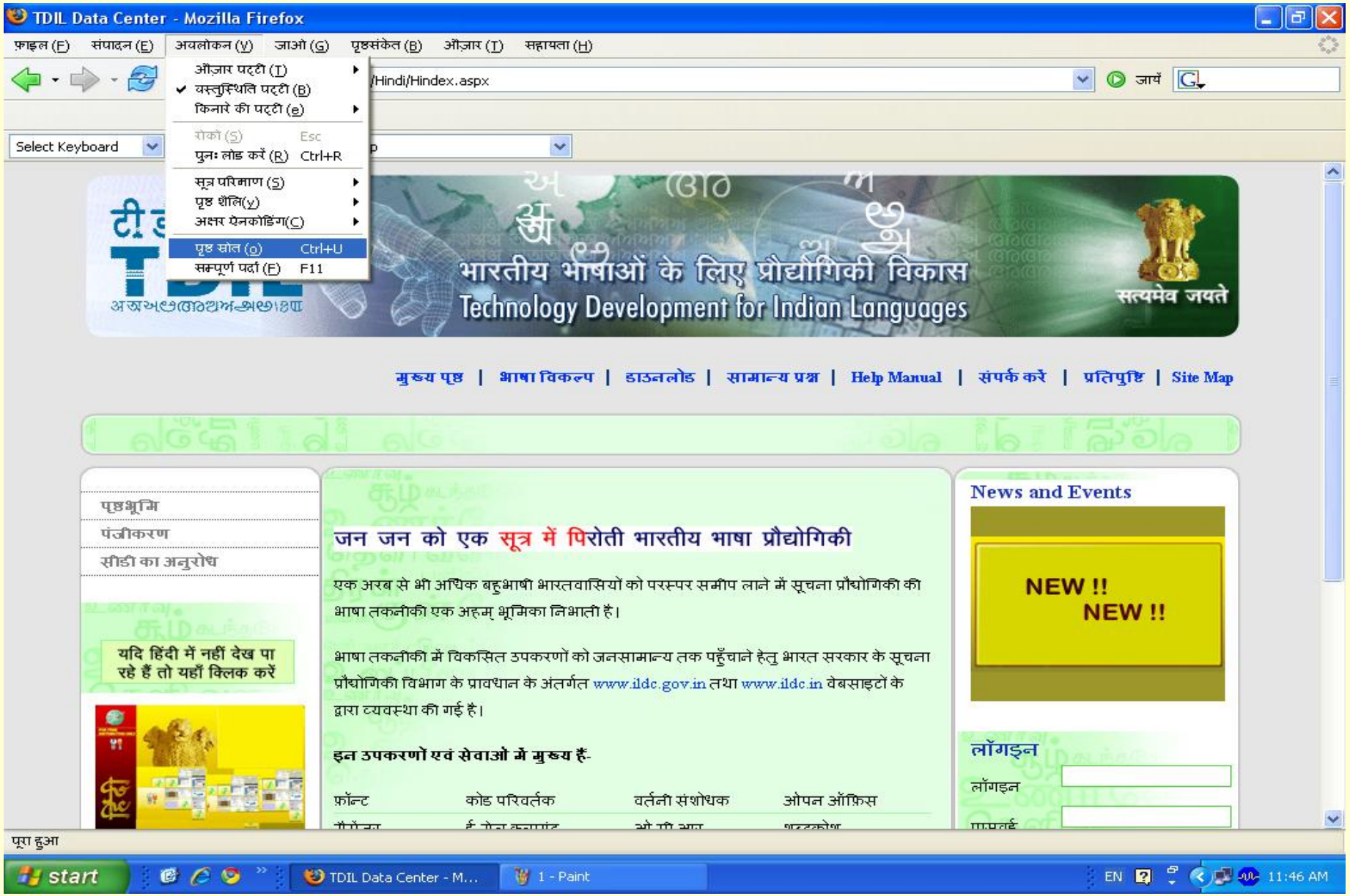
Text to Speech System with Braille Interface in six Indian Languages : Hindi , Tamil , Telugu , Bengali , Marathi and Malayalam Languages

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



Localized TDIL data Centre Portal in Mozilla Firefox



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, Guajarati 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

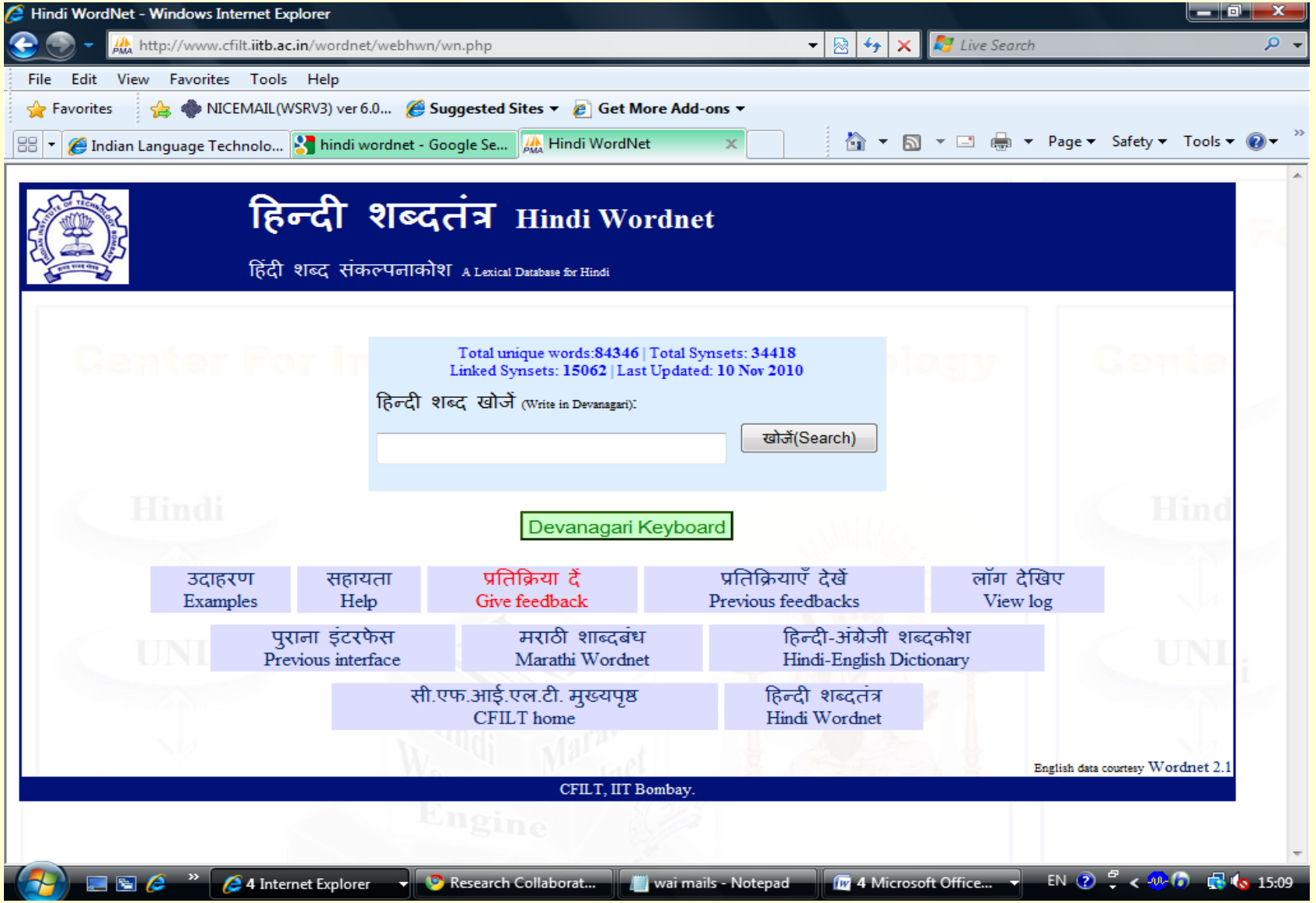
- Five levels of annotation namely phoneme, syllable, word, phrase and parts of speech (POS) are used for annotation.

- The Annotated speech signal and its output i.e. standard format time table (SFT) are available.

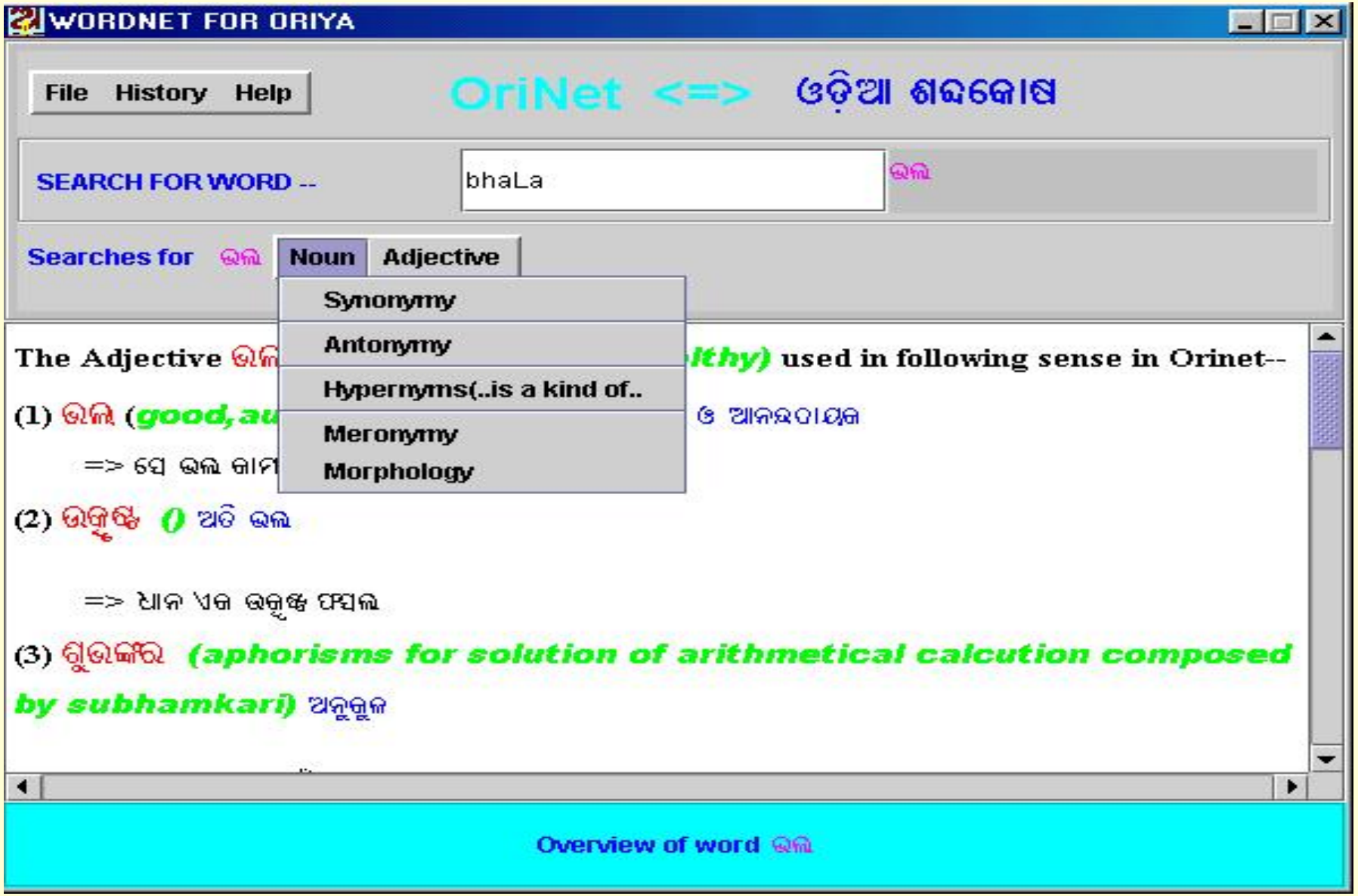
Linguistic resources development under Consortium Mode projects

- 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



Oriya Word-Net

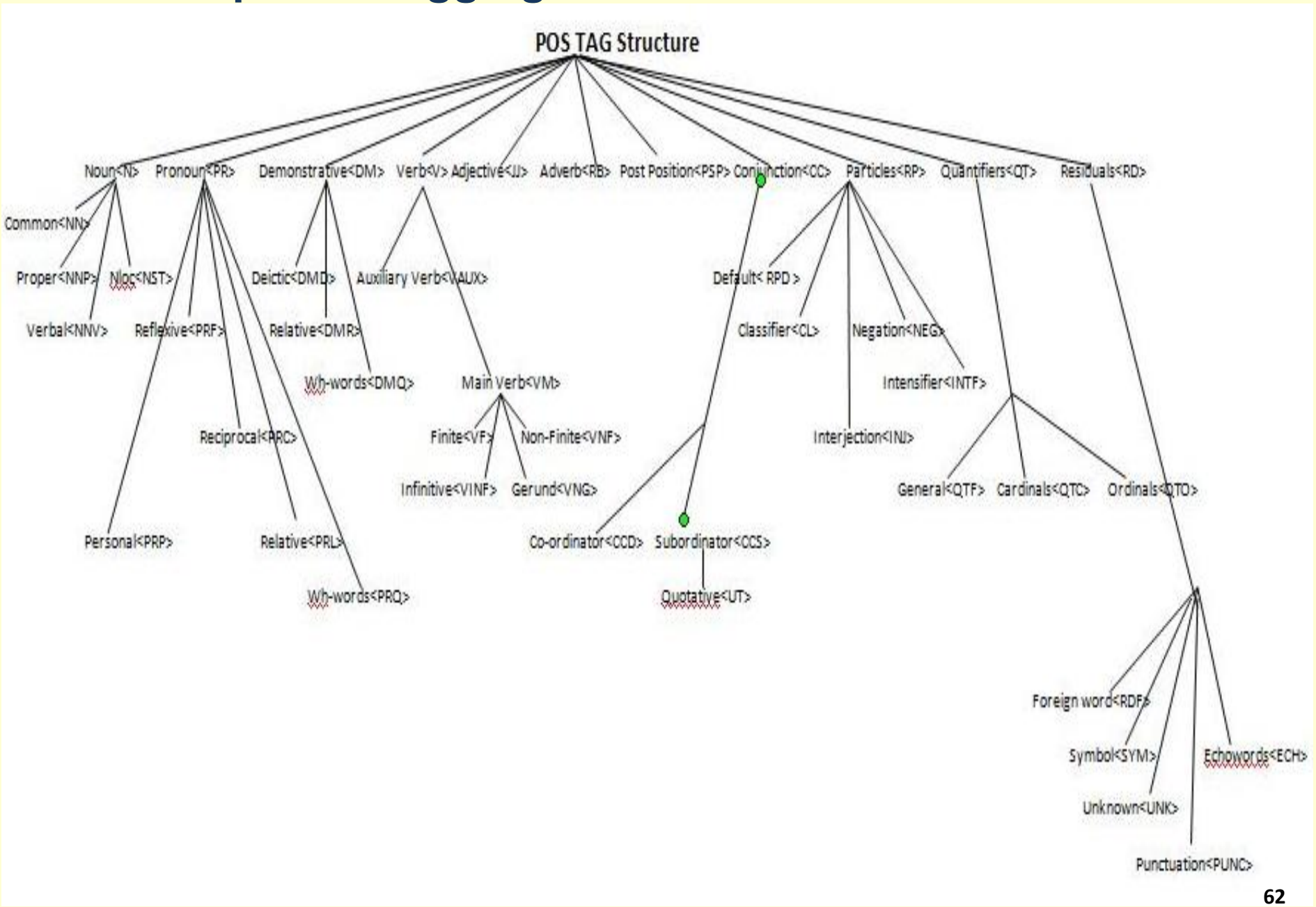


STANDARDIZATION EFFORTS

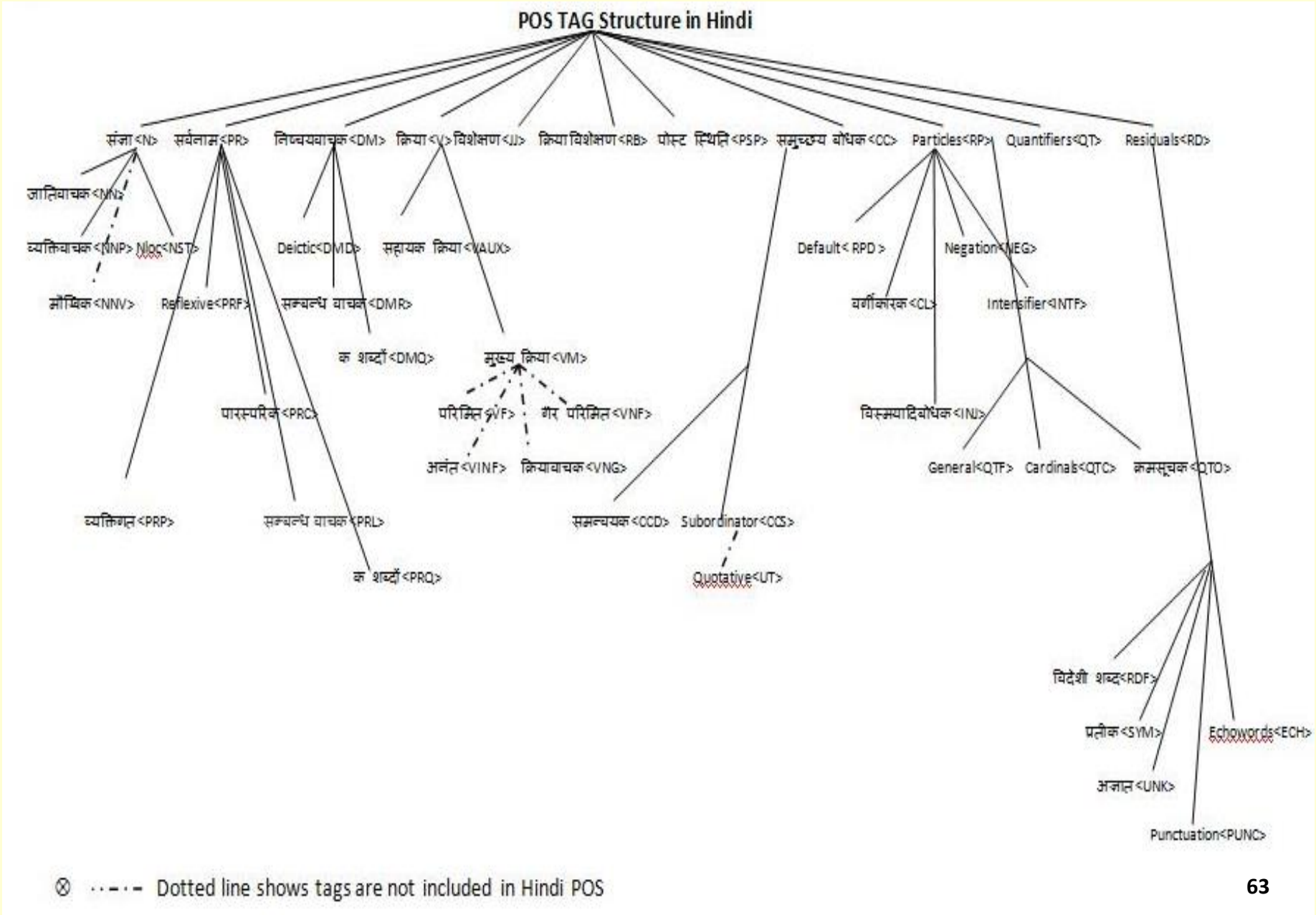
STANDARDIZATION EFFORTS

- **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



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 (equvalant to Shure, Sennheiser,etc.) **Preamp.: 30Hz-15kHz** Sound Card: Creative Gold
2	Recording Environment	Speech studio (SNR >= -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 *Note : the defination of the Phone, Syllable, Word boundry in contineous* *Speech as given in annotation guidele (Annexure-II)* **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.

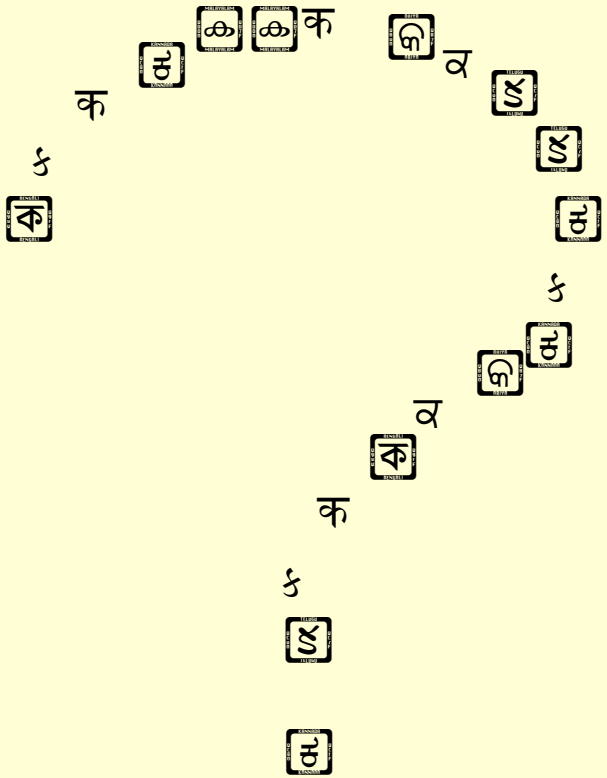
Snippet Generation:

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.**

Thanks & Questions



slata@mit.gov.in

☎ 91-11-24363535