Challenges of Multilingual Web in India : Technology Development & Standardization Perspective





Swaran Lata Country Manager , W3C India & Head , TDIL Programme , Dept of Information Technology , Govt of India E-mail : swaran@w3.org, slata@mit.gov.in



Organization of my talk :

- India Internet Scenario in India
 - Multilingual Web based service in India
 - Major web based services in India
 - : E-Governance Common Service Centers & Citizen centric services
 - : Unique ID Project Aaadhar

Multilingual Complexity In India:

- Origin of Indic Scripts & family
- Complex Characteristics

Technology Challenges

- Ab-initio development of Language Computing Solution
- Non-Conformance with English –Centric Model
- Complex nature of key technologies (Machine Translation, Cross-lingual Information Access, Character Recognition etc)
- Some key Initiatives

Standardization Challenges:

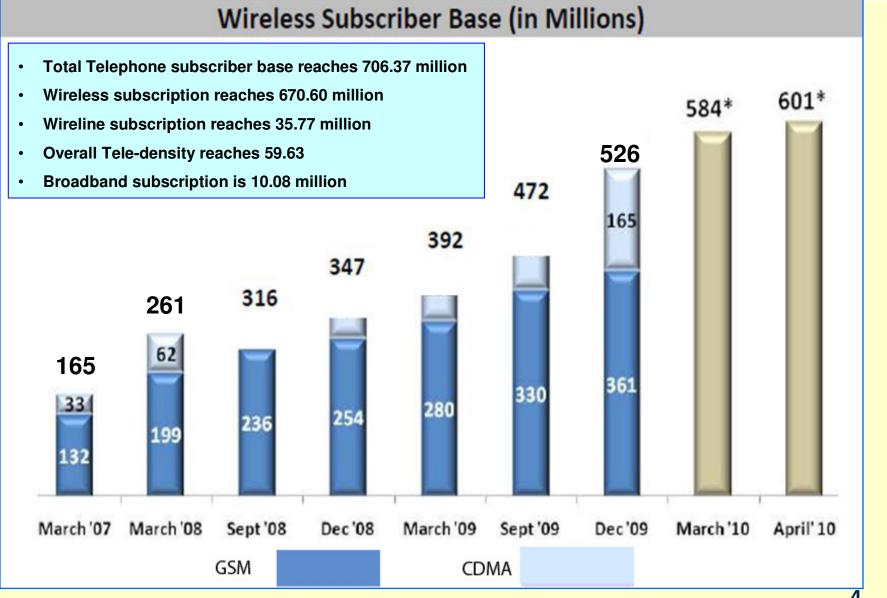
- Core :
 - -Encoding
 - -Input
 - -Display
- Middleware : HTML ,CSS , Mobile Web , Web Accessibility & Mobile Web
- Application Level : Presentation Issues & Web design implementation issues



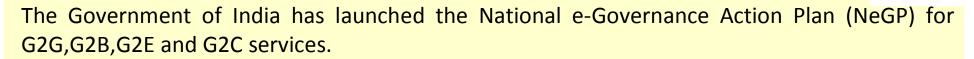
Internet Usage Statistics in India भारत →-%Pen. 4 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2009 2010 Year

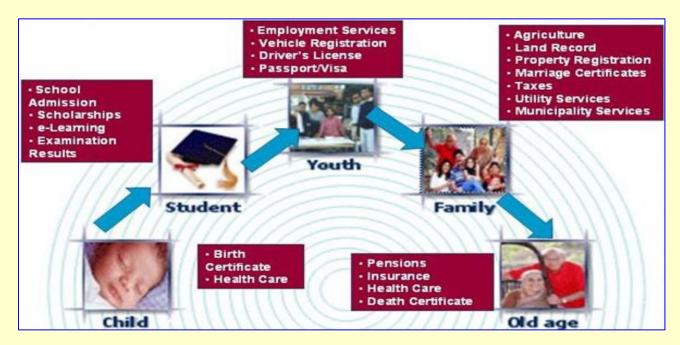
Wireless subscriber Statistics in India





E-Governance – Common Service Centers & Citizen centric services





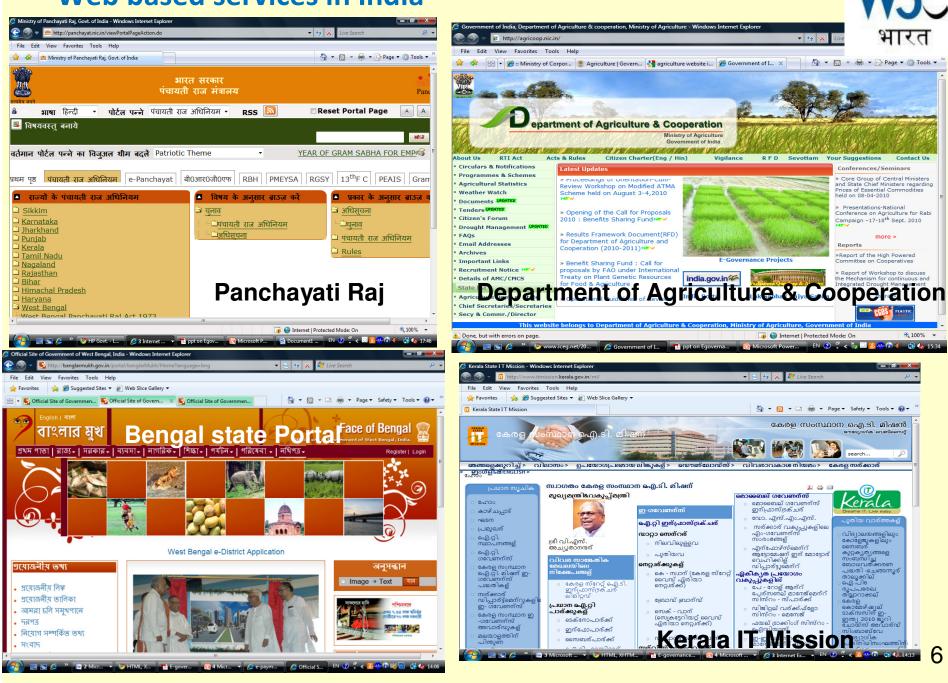
Modalities for Government on the Web

Provide: public services on the web, either transactional or information services or both.

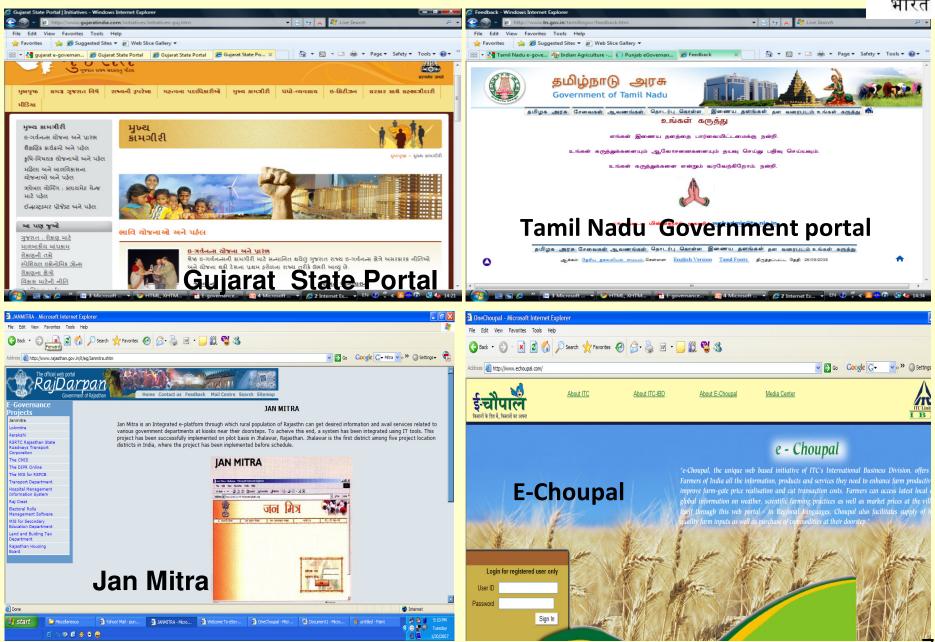
Engage: with citizens and businesses, on government terms or on the citizens terms.

Enable: public sector information re-use .

Web based services in India



Web based services in India



Online Services under National e-Governance Plan



Central Mision Mode Programme (MMP)s

- Income Tax
- Passport/VISA
- Company Affairs
- Central Excise
- Pensions
- Land Records
- Road Transport
- Property Registration
- Agriculture
- Municipalities
- Police
- Employment Exchange
- E-Courts

State Mision Mode Programme (MMP)s

- Agriculture
- Commercial Taxes
- e–District
- Employment Exchange
- Land Records
- Municipalities
- Gram Panchayats
- Police
- Road Transport
- Treasuries



Nation-wide Unique ID Project

• The Unique Identification Authority of India (UIDAI) was established in 2009 by the Government of India, with the developmental mandate of setting up the infrastructure to provide a universal way of uniquely identifying Indian residents.

ADHAAR

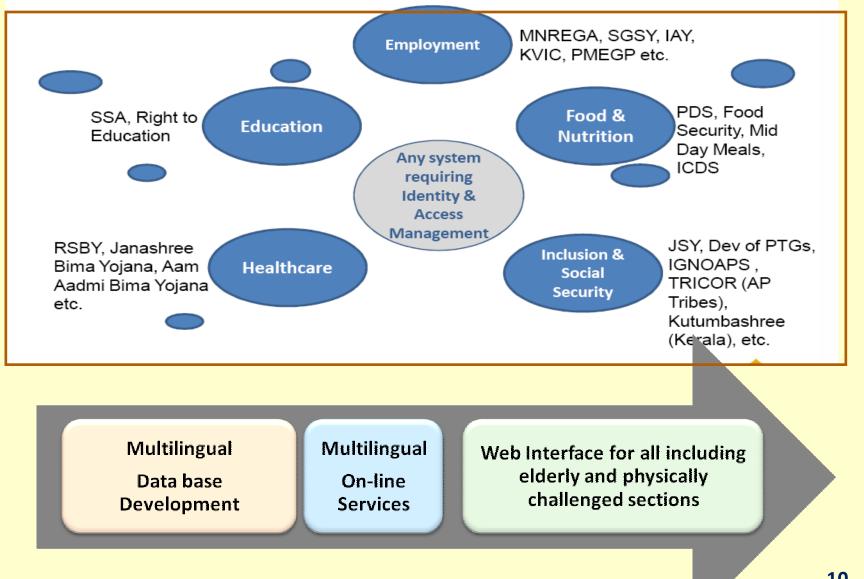
- AADHAAR, a 12-digit unique identification number (UID) that will be provided after getting the demographic and biometric information of an individual.
- AADHAAR's guarantee of uniqueness and centralized online identity for building the multiple services and applications
- AADHAAR give ability to access these web based services and resources, anytime, anywhere in the country

Aadhar – Towards foundation of Nation-wide multilingual web based services accessible to its citizens irrespective devices & platforms



Host of On-line Nation-wide Web Based Services for a population over 1 billion





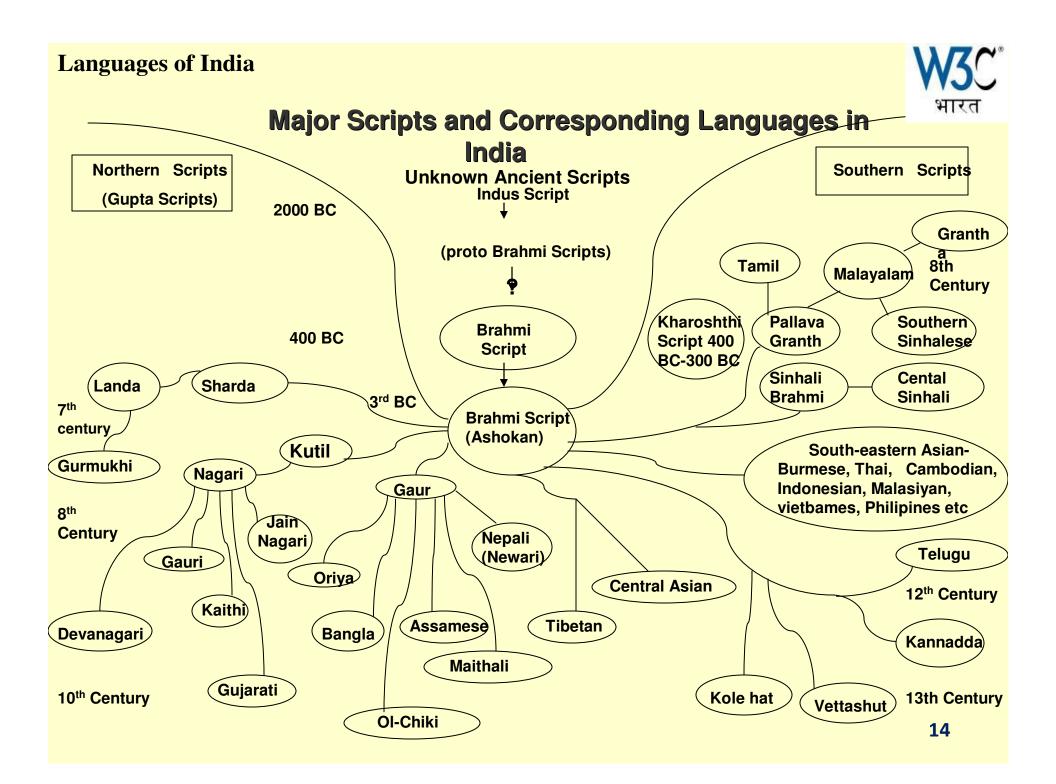
Multilingual Complexity In India

Languages of India



- According to Census 2001 India has 122 major languages and 2371 dialects.
- Out of 122 languages 22 are constitutionally recognized languages.
- Linguistic Diversity is very rich and wide in India
- One Language –many scripts
- Many Language –one script
- Culturally different depending on region though using same script for different languages.
- Even wide difference for same language across different parts of the country

		States							
						– भारत			
	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			
anguages	Goa	Gujarat	Haryana	Jharkhand	Lakshadweep	Meghalaya			
Bu	Konkani	Gujarati	Hindi	Hindi	Malayalam	English			
'an	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	Malayalam	Sanskrit	Bengali				
	Urdu	Dogri	Telugu	Urdu	Tamil, Telugu				



Technology Challenges

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



Complex nature of key technologies : Machine Translation

	anguages Machine Trans gram ,Department of Information Technology ,Gov Ho	· · · · · · · · · · · · · · · · · · ·
* This page can disp Input	lay maximum of 5 outputs at a time. Output	More
The best time to visit Jaipur is December to February	जयपुर दर्शन करने के लिए सर्वोतम समय वि से फरवरी है	देसंबर 💦 📋
Bara Shingri Glaciers is the largest in Himachal Pradesh	बारा शिंगरी ग्लेशियर हिमाचल प्रदेश में सबर है	से बड़ा 💦 📗
The Jal Mahal is a picturesque palace built for royal duck shooting parties	जल महल शाही बत्तख आखेट दल के लिए वि किया गया जीवन्त महल है	नेर्माण 💦
<		+
	Final Output	
IISc-Banglore,Utkal Universi	C Pune ,IIT Bombay , CDAC Mumbai , Jadhavpu ity,Banasthali Vidyapeeth ,Amrita University hbad , IIIT Hyderabad	ur University,

Sample Outputs For English - Urdu



Ukeer : Mohan Language: Urdu PrintTarget	Text PrintEnglishText RebuildTargetText Translation SignO
	al Output
The Amber fort complex has several apartments with palaces, halls, stairways, pillared pavilions, gardens & temples.	≞ ۵۔ امبعیر قلعہ میں محلوںہ دیوان خانوں ، سیڑھیوں ، ستونوں€ سے سچے گنبد ، باغوں اور مندروں جیسے کئی حصے ہیںک
Amber palace is a classic example of Mughal & Hindu architecture.	۶۔ امبیر محل مغل اور ـیندو معماری کا ہے مثال نمونہ ہے۔
Sa	ve in Inbox

Sample Outputs For English - Bangla



Iser: a Language: Bangla	
210	t1 argetText PrintEnglishText RebulidTargetText Translation SignOut Final Output
The Amber fort complex has several apartments with palaces, halls, stairways, pillared pavilions gardens & temples. Jaipur, popularly known as the Pink City, is the capital of Rajasthan state, India.	বাগান এবং অন্দিরসম বিভিন্ন স্বতয়মর আছে।
	হ হ

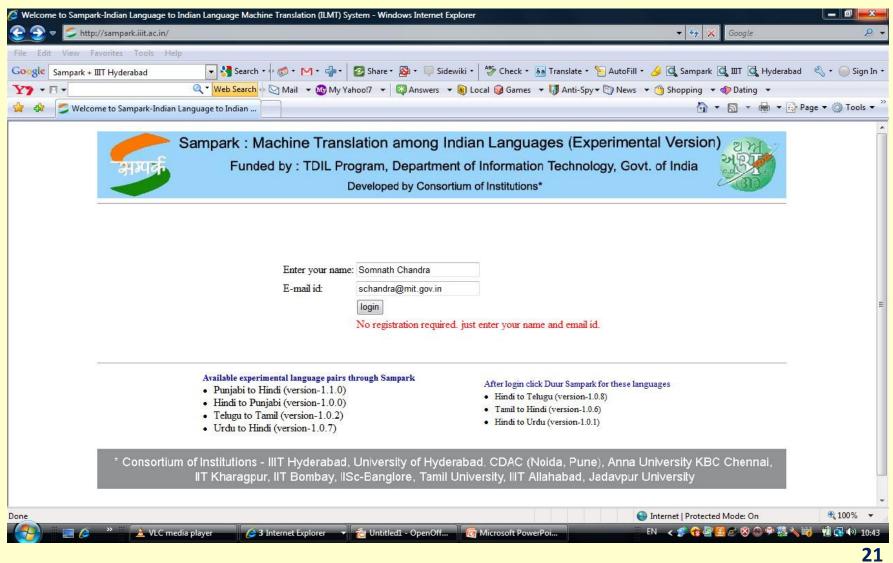


Sample Outputs For English - Tamil

Ukaar : Mohan Language: Tamil	
	getText PrintEnglishText RebuildTargetText Translation SignOut
F	final Output
ime to visit the town. Guru Shikhar is the highest peak on the Mount Abu which provides an excellent view of the whole town.	ஏப்ரல் முதல் ஜீன் மற்றும் அக்டோபர் முதல் நவம்பர் வரையுள்ள காலங்கள் நகரைச் சுற்றிப் பார்க்க சிறப்பான காலங்கள், மவுண்ட் அபுவின் மிக உயர்ந்த சிகரமான குருசிக்கார் நகரத்தின் சிறப்பான காட்சியினை முழுமையாகத் தருகின்றது.
	Save in Inbox



Indian Language to Indian Languages Machine Translation System



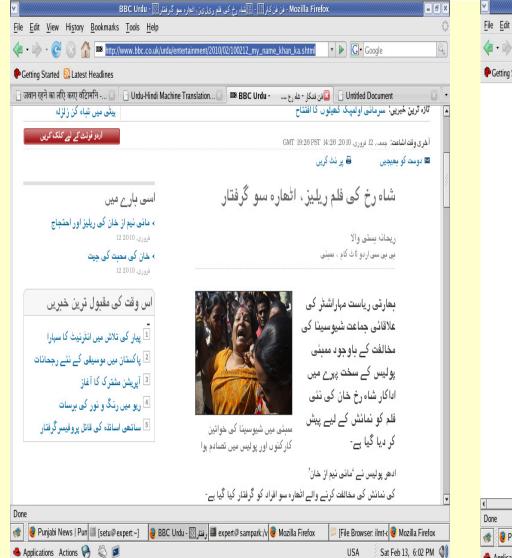
Sample Outputs For Punjabi - Hindi



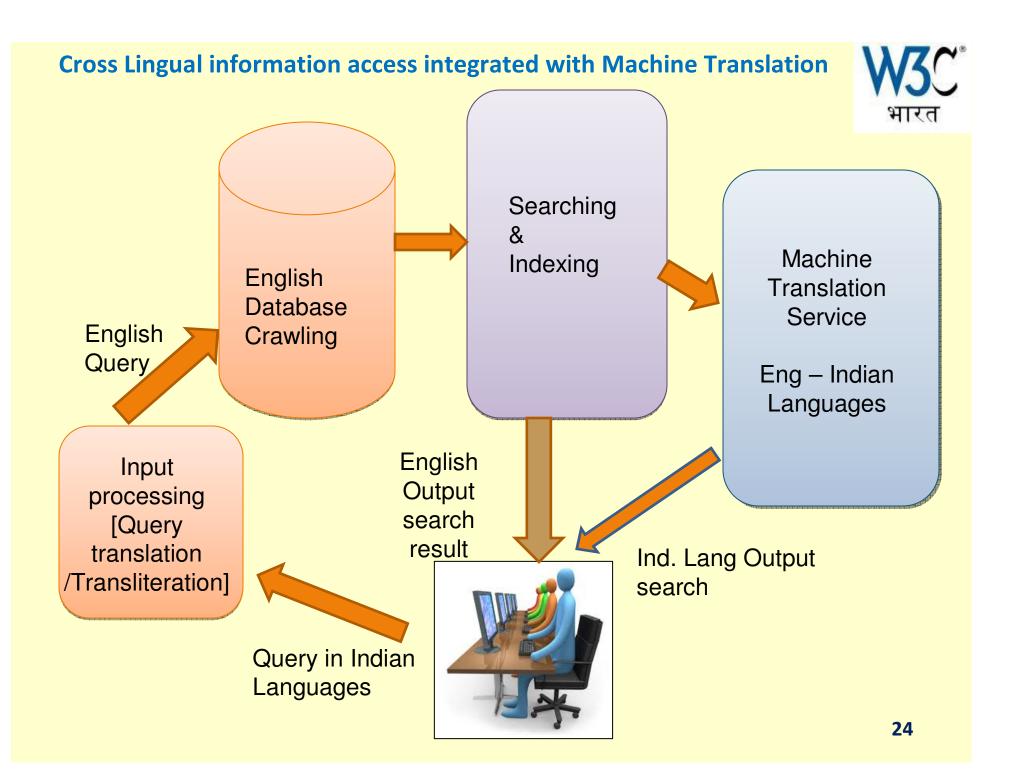
🖉 Welcome-Indian Language Machine Translation (Sampark) - Windows Internet Explorer								
G v fttp://sampark.iiit.ac.in/sampark/src/phpscript/translate_text.php	 ✓ 4 × Live Search 							
Retrieving buttons from Ask.com								
💽! 👻 🔍 🔍 Web Search 🕂 🖓 Mail 👻 🐼 My Ya	shoo! 👻 🖾 Answers 👻 🔰 Anti Spy 🛛 👰 Bookmarks 👻 🗁 Messenger 🕶 🏨 Movies 👻 🔷 👋							
😪 🎄 🔠 🕶 🍘 ICICI direct.com :: Tra 🙋 NICEMAIL(WSRV) ver 📁 Welcome-	Indian L X 👿 ਪੰਜਾਬੇ ਕੋਗੇ - ਵਿਕਿਪੀਡਆ 🍈 🔻 🗟 🔻 🖶 👻 Page 💌 🎯 Tools 👻							
Funded by : TDIL Program, Depart	g Indian Languages (Experimental Version)							
Enter input text for translation in Unicode (upto 180 words)	Translation: Punjabi-Hindi You are: Manoj Logout							
ਪੰਜਾਬੀ (ਜਿਸ ਨੂੰ ਅੰਗਰੇਜ਼ੀ ਵਿੱਚ Punjabi, ਗੁਰਮੁਖੀ ਵਿੱਚ ਪੰਜਾਬੀ ਅਤੇ, ਸ਼ਾਹਮੁਖੀ ਵਿੱਚ ਸ਼ੱਦੀ ਦਾਂਗ ਲਿਖਿਆ ਜਾਦਾ ਹੈ।) ਭਾਰਤ ਅਤੇ ਪਾਕਿਸਤਾਨ ਦੇ ਪੰਜਾਬ ਸੂਬੇ ਦੀ ਭਾਸ਼ਾ ਹੈ। ਇਹ ਭਾਰਤੀ- ਇਰਾਨੀ ਵਰਗ ਦੇ ਵਿੱਚੋਂ ਭਾਰਤੀ-ਯੂਰਪ ਵਰਗ ਨਾਲ ਸਬੰਧਤ ਹੈ। ਇਸ ਤੋਂ ਇਲਾਵਾ ਪੰਜਾਬੀ ਸ਼ਬਦ ਨੂੰ ਪੰਜਾਬ ਨਾਲ ਸਬੰਧਤ ਕਿਸੇ ਵੀ ਚੀਜ਼ ਲਈ ਵਰਤਿਆ ਜਾਂਦਾ ਹੈ, ਜੋ ਕਿ ਪੰਜਾਬ ਜਾਂ ਪੰਜਾਬੀ ਨਾਲ ਸਬੰਧਤ ਹੋਵੇ, ਜਿਵੇਂ ਕਿ ਪੰਜਾਬੀ ਬੋਲਣ ਵਾਲਿਆਂ ਨੂੰ ਪੰਜਾਬੀ ਅਤੇ ਪੰਜਾਬੀ ਖੇਤਰ ਵਿੱਚ ਪੰਜਾਬੀ ਹੀ ਕਿਹਾ ਜਾਦਾ ਹੈ।	पंजाबी जिस को अंग्रेजी में फुन्जबि , गुर्मुखी में पंजाबी एवं , शाहमुखी में لنجابی का माफ़िक लिखा ???? है । भारत एवं पाकिस्तान के पंजाब सूबे की भाषा है । यह भारतीय - इरानी वर्ग में भारतीय - यूरोप वर्ग से सबंधत है । इस से अलावा पंजाबी शब्द को पंजाब से सबंधत किसी भी चीज के लिए बरता जाता है , जो कि							
Type in Choose	Translate Select Language Translate							
Translate a web page (Give URL below)	Select Language 🔻							
Inttp:// Select Language • Translate Help Feedback Sample text Sample URL Download plugin Translate Document New! Duur Sampark New! * Consortium of Institutions - IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University KBC Chennai,								
III Kharagpur, III Bombay, IISc-Banglore, Ta	mil University, IIIT Allahabad, Jadavpur University							
A http://in.movies.vahoo.com/								
✓ ////////////////////////////////////	III Protected Mode: On Calutor of the Calutor of th							

Urdu-Hindi ILMT Translation through URL :- Urdu input

(Source web page : http://www.bbc.co.uk/urdu/entertainment/2010/02/100212_my_name_khan_ka.shtml)







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.

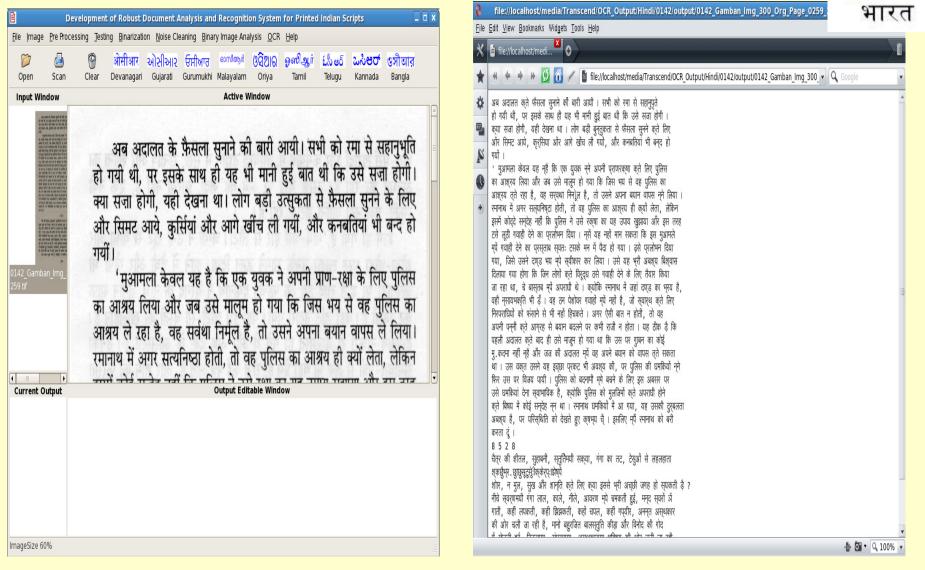


মার্টা search কলকাভা	<u>উন্নত থেঁজ</u> সাহায্য কী-বোর্ড	VV3
	<u>কা-বোড</u>	भारत
(খাঁজ		
থোঁজ : 💿 এই ইঞ্জিন দ্বারা 🔘 অন্যান্য ইঞ্জিন	ারা	
হিন্দি ইংরাজী বাংলা	পরিণাম 1 খেকে 10 মোট 1221 থোঁজ -	
কলকাতা যদিও ভারতের অন্যান্য মহানগরগুলির মতো নগরায়ণ সমস্যার অঙ্গ হিসাবে দারিড়, পরিবেশ দূষণ ও যানজটের সমস্যা (http://bn.efactory.pl/কলকাতা (<u>ভান্ডার)</u> (<u>সারাংশ</u>)	ক একেবারে মুক্ত যডে পারেনি কলকান্ডা। কলকান্ডা শহরের প্রসিদ্ধি তার বৈপ্লবিক আন্দোলনগুলির জন্য।	
কলকাতা মেট্রো		
কলকাভার বিভিন্ন খানবাহন সম্পর্কে বহু অভিযোগ থাকলেও মেটো সম্পর্কে অভিযোগ থুবই কম। কলকাভার মেটো সা	Whttp://bn.wikipedia.org/wiki/%E0%A6%95%E0%A6%95%E0%A6%95%E0%A6%95%E0%A6%BE%E0%A6%AE%E0%A6%AE	🔻 🔯 🐓 🗙 🚼 Google
http://bn.efactory.pl/কলকাতার_(মট্রো (<u>ডান্ডার</u>) (<u>সারাংশ</u>)	🗙 🎽 AVG 👻 💽 🕶 🐨 👻 😌 🐨 🐨 👻 😴 Search 🔶 🚷 Total Protection 🛛 🚳 AVG Info 🝷 Get More 📄 🤕 - 🔫 Identity Guard 🔶 😭 Favorites 🛛 🎭 🔽 New king of tech Apple o 🏠 Printable Summer Them 👔 Summer Coloring Pages 2 📚 Printable Summer Colorin 🖤 Lecture	e Notes 👩 Suggested Sites 🔻 💋 Web Slice Gallery 👻
কলকাভা মেট্রো	🔐 🔹 🕅 India Result Page 🥖 स्वरूता: 🗰 🐨 🗰 🗰 🕅 🗰 🕅 🗰 🕅 🗰 🕅	🛐 🔻 🔊 👻 🖶 🕈 Page 🕶 Safety 🕶 Too
কলকান্ডার বিভিন্ন মানবাহন সম্পর্কে বহু অভিযোগ থাকলেও মেটো সম্পর্কে অভিযোগ থুবই কম। কলকান্ডার মেটো সা http://bn.efactory.pl/কলকাতা_(মটো (<u>তান্ডার</u>) (<u>সারাংশ</u>)	উইকিপিডিয়া ত How to read and write in Unicode Bangla 🛕 Must-read for	contributors
<u>কলকাতার ইতিহাস</u> Searching for: কলকাতার ইতিযাস ? কলকাতা সর্ববৃহৎ শহর ও প্রধান বন্দর হিসেবে বিরাজমান এ মহানগরীটি ১৯১	একটি মুদ্ত বিশকোষ কলকাতা প্রধান পাতা	
http://bn.efactory.pl/কলকাতার_ইতিহাস (<u>ডান্ডার</u>)	সম্প্রদানের প্রবেশহার "কবিকাতা" এখানে পুননির্দেশ করা হয়েছে। অন্য ব্যবহারের অন্য, দেখুন কবিকাতা (দ্রার্খতা নিরসন)	Ten things you may not know about Wikipedia
<u>কলকাতা - উইকিপিডিয়া</u> যদিও ভারতের অন্যান্য মহানগরগুলির মতো নগরায়ণ সমস্যার অঙ্গ হিসাবে দারিদ্র, পরিবেশ দূষণ ও যানজটের সমস্যা (http://bn.wikipedia.org/wiki/কলকাতা (<u>ভান্ডার) (সারাংশ)</u>	সমসাদরিক ঘটনা ক লকাভা (পূর্বনাম: কণিকাভা ; ইংরেজি: Kolkata, পূর্বে 4) Calcutta ^{আকাত্রত} া তারডের পূর্বাঞ্চণীয় রাজ্য পশ্চিমবঙ্গের রাজঘানী, প্রধান বাণিজ্যকে সাম্প্রতিক গরিবর্তনসমূহ অহানা খেকোনো গৃষ্ঠা বৃহত্তর কলকাভার অনসংখ্যা ১ কোটি ৪০ শক্ষের কাহাকাছি। এই অনসংখ্যা ৫০ লক্ষের কিছু বেশি। অবে কলকাভা ও ভার পার্যবর্জী জেলাগুলির অংশা অহানা খেকোনো গৃষ্ঠা বৃহত্তর কলকাভার অনসংখ্যা ১ কোটি ৪০ লক্ষের কাহাকাছি। এই অনসংখ্যার বিচারে কলকাভা তারডের চন্ডুর্থ বৃহত্তম শহর ও ভূতিায় বৃহত্তম মেট্রাপদিট সাযান্যা অঞ্চল এবং বিশ্লের অষ্টম বৃহত্তম মহানগর অঞ্চল। ¹⁰ কলকাভা পৌরগ্রপাচার উত্তর চিবিশ পরগনা, পূর্বে উত্তর ও দহিগ চরিশে পরগনা এব দান করুন্দ দরিগ চরিশ পরগনা জেনা অবস্থিত। পশ্চিম দিকে হুগনি নদী এই শহরেক হাওজা জেলা থেকে বিশ্বিদ্ধ র রেছে।	বিশেষ নিয়ে গঠিত টান বা মহানগরীয়
	 মূডগ/এঙ্গাণোট ১৭৭২ সালে মূর্শিদাবাদ শহর (থকে বাংদার রাজধানী কলকাভাস্ত স্থানান্তরিত করা হম। ১৯১১ সাল পর্যন্ত কলকাভা শুমুদ্র বাংদারই নয়, বরং সমগ্র রি বই ভৈরি কঙ্গল রাজধানী ছিল। ১৯২৬ সালে ক্যালকাটা মিউনিসিগ্যাল আক্টের অধীনে কলকাভার স্থানীয় স্বায়ন্তশ্যাসন কর্তৃপক্ষ কলকাভা পৌরসংস্থা স্থাপিত হম। ১৯৪৭ PDF ডাউনলোড ঘণার যোগ্য মংদ্ররণ পীঠস্থান। ১৯৫৪ সালের পর পেকভান বগঠিত পশ্চিমবন্স রাজের রাজধানী যোশিত হয। এই সময় কলকাভা ছিল আধুনিক ভারভের শিক্ষা, বিজ্ঞান, শিব, সংস্কৃতি গে ঘণারা হার্যান বিজ্ঞান, বিদ্যাল সংস্কৃত্য কলকাভা নবগঠিত পশ্চিমবন্স রাজ্যের রাজধানী যোশিত হয়। এই সময় কলকাভা ছিল আধুনিক ভারভের শিক্ষা, বিজ্ঞান, শিব, সংস্কৃতি গে ঘণার যোগ্য মংদ্ররণ পীঠস্থান। ১৯৫৪ সালের পর থেকে রাজনৈত্তিক অস্থিনতা ও অর্থনৈত্তিক অবহারের ফলে সেই গৌরব আনকাংশে শ্ব হয়। ভবে ২০০০ সালের পর থেকে আরনেতি বন্ধির হার এর সাংস্কৃতিক ভারতারিব আনকাংশে শ্বনংছনা র করে। বাণিও তারভের আন্যান শহরের মত্যে কল 	। সালে ভারত ও রাজনীতির এক এই মহর পুনরায়
Bengali Monolingual Retriev	নগানা ভাবাসমূহ কাগানা ভাবাসমূহ Afrikaans Afrikaans Aragonés Englisc এই শহর ৫ বিশেষ ভূমিকা গ্রহণ করেছিল। অন্যদিকে আধুনিক ভারভের প্রধান প্রধান সংস্কৃতিক আন্দোলনগুলিরও প্রাণকেন্দ্র ১ই কলকাভা। এই কারণে Aragonés Englisc এই শহরকে <i>আকল নগরী</i> বা, সিটি অফ লব নামেও অভিষ্ঠিত করা হয়। রাজা রামমোহল রায়, রবীন্দ্রলা বার্টপুর, স্বামী বিবেকানন্দ, রোনান্ড রম, সূতা এ, এ এ, এ এ, এ এ, এ এ, এ এ, এ এ, এ এ,	াণবালোলনগুলিতে গ এই শহরকে গহাবহানের অন্য ষচন্দ্র বসু, মাণার

Best Turne Visc Caku. Cakuta Fairfeetual. Cakuta Fairfeet		start=10&hitsPerPage=10&sessionlang=bn⟨=en 👌 Total Protection 🔊 AVG Info • Get More 🤕 • 🞯 Identity Guar em 🖉 Summer Coloring Pages 2 🍨 Printable Summer Colorin 🎙		W3C * भारत
Calcuta Mureums and Monument, Indam Mureum, Calcuta City Guide, Calcutative and participants and participant	শোঁজ: এই ইঞ্জিন দ্ব হিন্দি ইংরাজী বাংলা	থোঁজ ারা © অন্যান্য ইঞ্জিন দ্বারা	পরিণাম 11 থেকে 20 মোট 35 থোঁজ	
Calcutta Travel Packages, calcutta Sports, calcutta India Tours and calcutta Travel Packages, calcutta Sports, calcutta India Tours and Travel, calcutta physing sports Kolkata festivals Kolkata festivals	Calcutta Museums and Monuments,Indian Museum,Calcutta City Guide,Calcutta located in Chowringhee on the কলকাতা মিউসিয়াম এবং স্মৃতিস্তস্থ , ইন্ডিয়ান মিউজিয়াম , কোলকাতা শহর f করেছিলেন , সবচেয়ে পুরানো একটি এটা এশিয়া জাদুঘরের মধ্যে উপর চৌরিঙ্গী	a Solution of tech Apple o Preschool Summer Them @ Summer Street S	+ 💽 🤣 🗙 Itection 🎳 AVG Info + Get More 🥳 + 🔫 Identity Guard ummer Coloring Pages 2 🌎 Printable Summer Colorin 🖤 Lecture Notes 🔊 Sugg	ested Sites 👻 🙋 Web Slice Gallery 🔹
Kali Puia	calcutta Travel Packages,calcutta Sports,calcutta India Tours and calcutta Travel Packages,calcutta Sports,calcutta India Tours and Travel,calcutta playing sports কলকাতা ব্রমণ প্যাকেজ, কলকাতা , কলকাতা , তারতের ব্রমণের এবং ব্রমণ , হয় http://www.calcutta-travel-packages.com/kolkata-sports.htm (তান্ডার) (সারাংশ) How to Reach,calcutta city guide,calcutta Travel Packages,calcut How to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute Mow to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute Mow to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute Mow to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute Mow to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages,calcute How to Reach,calcutta city guide,calcutta Travel Packages, calcute Mitting State , former on the state Mitting State , former on the state Mitting State	Image:	weak Kolkata festivals hamii) Tourist attractions wealth, peace & prosperity Calcutta Religious places e month of October, about 5 days after ta India is an important occasion, in which Goddess of wealth, peace and prosperity. The ty to people to invite the Goddess of luck and wo days before the celebration of the festivity Kolkata are beautifully decorated. The shops eated on a lotus. Museums in Kolkata y February s of learning Travel tips Kolkata sports in Kolkata that takes place during the period ebruary. In fact, the day when puja of Ma ate holiday. Calcutta Saraswati puja festival is w. It is on this day that the youngest female yellow dothing. Sarswati puja of Kolkata India ng. Hotels ost every locality of Calcutta. People of the t the pundal to celebrate the festivity. Kids are t is during this puja that children pray to the it is usually the pundit who performs the puja. Foreign embassies	1 - La Kange v Safety v

India Result Page - Windows Intern					Ŀ	
	.ac.in/clia-latest/result.jsp?TextArea=কলকাজ্যsessionlang=bn&la			🕶 🗟 🐓 🗙 🚼 Google		CVV
× Harrie Q!• बनवाडा	👻 💽 Search 🔶 🕤 Total Protection tech Apple o 🌈 Preschool Summer Them 🔊 Summer (-	🛛 🤕 🔹 👰 Identity Guard	Notor 🔊 Suggested Sites =	Web Slice College -	भारत
			-		✓ Page ▼ Safety ▼ To	-
😬 🔹 📈 India Result Page 🛛 🗙	🥖 মনসন্না: 🗰 W কাকাতা - উইকিপিউরা	Best Time to Visit Calcu.	<i> (C</i> alcutta Fair/Festival,Ca			
TINDLA search	লকাভা		উন্নত থোঁজ <u>সাহায্য</u> কী-বোর্ড			
	থোঁজ (খাঁজ : এই ইঞ্জিন দ্বারা 	ল দ্বারা				
হিন্দি ইংরাজী বাংলা				পরিণাম 1 (থকে 10 মোট 348 থোঁজ	
कलकत्ताः इसका पूर्व नाम अंग्रेजी में भले ही "कैलक এর পূর্বে नाम देংরেজ সালে যদিও যে পরিচিত ছিল http://hi.pandapedia.com/wiki/कलकत्त	cr हो लेकिन बंगाल और बांग्ला में इसे हमेशा से कोलकाता या कोलिक 51 (केनकाछा ' হয়ে किन्छ বাংলার এবং বাध्श्रा সালে এটিকে সব দ জেলাব) (আবদ্য)	সময × 🛔 AVG 🔻 🞯!- লালারা	ndapedia.com/wiki/%E0%A4%95%E0%A4%B2%t	🗸 👌 Total Protection 🛛 🖏 AVG Info 🗸	Get More 📔 🤕 👻 📵 Identity Guard	Cure Notes Suggested Sites Web Slice Gallery
пар./па.рапаареша.соп/ wiki/чесчес		吕 🔻 🌈 कलकत्ताः	× 🏉 कलकता: 🛛 🕅 वगवा	ল - ভইকিপিডিয়া 🏉 Best Time to V	'isit Calcu 🌈 Calcutta Fair/Festival,C	a 👌 🔻 🖾 👻 🚍 🖶 👻 Page 👻 Safety 🕶 To
Dhapa কলকাতা দ্য়া (কন্দ্র তা : লাগ http://www.dhapa.com/hi/calcutta-me	महरा घर के बारे में भेजें एक पोस्ट Dhapa कोलकाता में चीनी समुदाय न प्रतूज वाफ़ी प्रम्थार्क (ভज़िन এकोर्टि (भाष्ठ Dhapa कनकाजा? rrcy-centre/ (<u>जाज़ात्र)</u> otepad, संस्कृति, स्मारकों और तस्वीरें	• <u>कोलकाता</u> • <u>कोलकाता</u> • <u>कोलकाता</u> • <u>Calcutta</u> • <u>Calcutta</u> • <u>Calcutta</u> • <u>Calcutta</u> • <u>GâcqC</u>	नाम कलकता) पश्चिम बंगाल की राजधा	<u>22.5697, 88.3697</u> बंगाल की खाडी व नी है। यह भारत का दसरा सबसे बड	ा महानगर तथा पाँचवा सबसे बडा ब	<mark>गली नदी</mark> के बायें किनारे पर स्थित कोलकाता (<u>बांग्ला:</u> र्ग्दरगाह है। यहाँ की जनसंख्या १,३२,१६,५४६ है। ¹¹ इस इ
भारत में दस शहरों के लिए मौसम का पूर्व छात्र (छत ५० भरतत जन) जनवा मु ख http://india.ine.cl/bi/2000/01/15/poper	লুमাল और अच्छे छुट्टी सेवा मतलब मई:) यह कलकता में. कलकत পৈডোগ পুর্বালুমান এবং ছুটির পরিষেবা অর্থ (ম :) এই কলকা		गाधाएँ अपने आप में समेटे हुये है। शहर द <u>साम्यवाद</u> आंदोलन के गढ़ के रूप में भी म अपनी उत्तम स्थिति के कारण कोलकाता यातायात का केन्द्र, विस्तत बाजार वितर	को जहाँ भारत के शैक्षिक एवं सांस्कृति 1ान्यता प्राप्त है। महलों के इस शहर व को पूर्वी भारत का प्रवेश द्वार कहा जा ण केन्द्र. शिक्षा केन्द्र. औद्योगिक केन	तेक परिवर्तनों के प्रारम्भिक केन्द्र 1 के सिटी ऑफ़ जॉय के नाम से भी ज ता है। यह रेलमार्गों, वायुमार्गों तथा द्र तथा व्यापार का केन्द्र है। अजायब	सड़क मार्गों द्वारा देश के विभिन्न भागों से जुड़ा हुआ है। र म्वर. चिडियाखाना, बिरला तारमडल, हावडा पल, कालीघ
Bengali	–Hindi Cross-lingu Retrieval	• <u>Calcutta</u> • <u>Калькутта</u> • <u>कलकोता</u>	विलियम, विक्टोरिया मेमोरियल, विज्ञान	नगरी आदि मुख्य दर्शनीय स्थान हैं यार करने का कारखाना, सूती-वस्त्र उ धत हैं। पूर्वांचल एवं सम्पूर्ण भारतवर्ष भ <u>ुमिका</u>	। कोलकाता के निकट हुगली नदी के चोग, कागज-उद्योग, विभिन्न प्रकार	: दोनों किनारों पर भारतवर्ष के प्रायः अधिकांश जूट के क के इंजीनियरिंग उद्योग, जूता तैयार करने का कारखाना, में कोलकाता का महत्त्व अधिक है।
						28

OCR

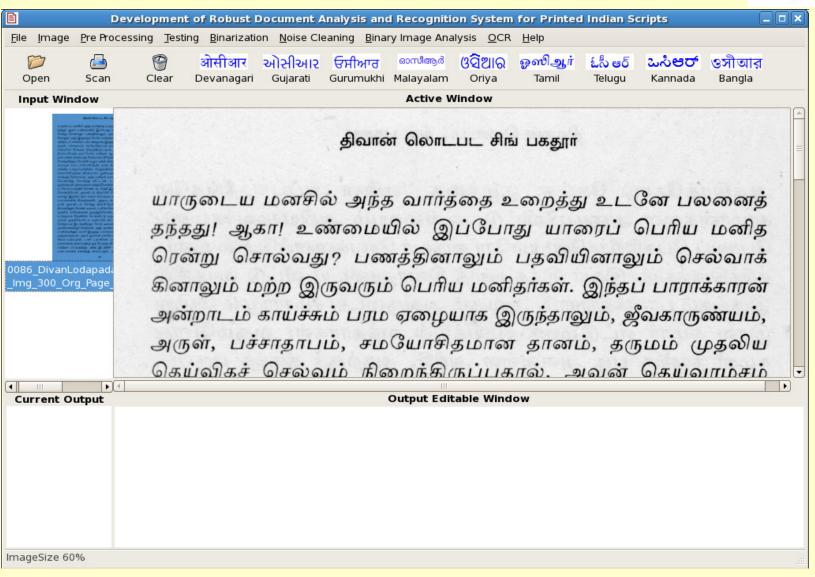


Original Image Loaded in Input Window of OCR

Output of Devnagari OCR

OCR



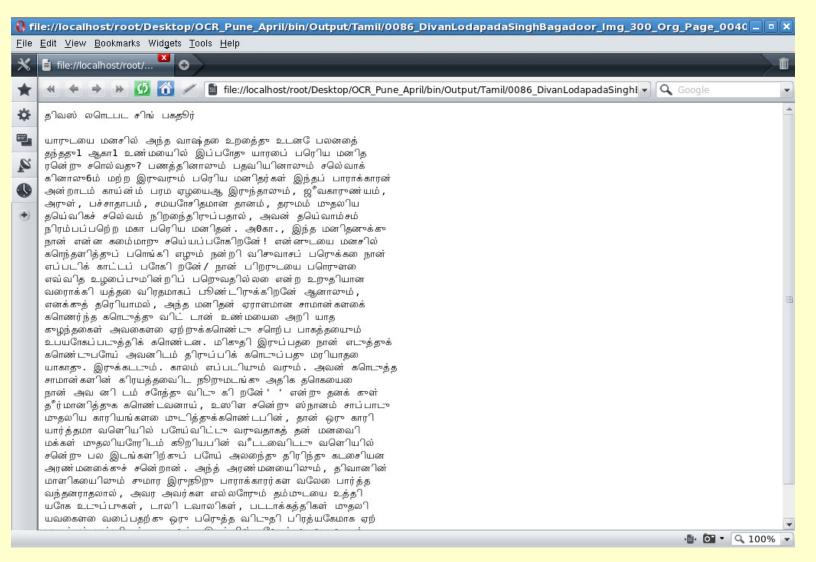


Original Image Loaded in Input Window of OCR

30

OCR





Output of Tamil OCR

Sample Tamil OHWR form - Tamil



🕒 😔 ĐI 🔚	XBA	🖋 🛛 X-Shij	ja 🔲	X-Scale	Y-Shift 1	S-See	le <mark>0</mark>	-	C50.top		
விமத்தல்கு உரினாளவல	tayınlar tiğiren sişişi e Dar Gundaya yaşı		ang	ш 1		uigung saint	rm cuainų ikar digra genr ču elegen		ஆம்மாண்ச் அ	jukca s udanta	1
	uster untur 🐴	जन्मे, 274	ы				ugi qi mnin	fer.endr.a.	.046~9. Den		
1. and a strate	2.000	jainiaj			1 andro or min		2	2. ແຫຼ່ງ ເກັກຈ			
@ 0 660 2741	1 4	H.W. 48			இப்கோ	874#		ச.ம.டசே	•0		
lajespags∟gerege angel orgå agan 225 07 2		6 ј ићс6 7 3.0	(ख्रु) ann+) 	(Gen) umadi	รัญหมู่คุ 5-1 อ อย่ 36	by orga D7	2009	а <u>нуте</u> 07	<u>சுல்க</u>] 30	leef) exect 7	
kangan sudang ut		_			4. وراية ماريخ	engi di	L00L0				
Signi .	URLA	\$ r.	พา คป		migarif		Ligoria		6.(FG84)		
ManLunni	லேக்கரி	3 6	கில்		ameur	nt	Cuteff		-প্রত জীন		
5.alemy.uldnaces					th الايراطة a.c.	ram e					
quinueri Gu Bit a	,				gimuri	Cungger C					
8. का जोन कु की जोन प्रमुखेल का जोन साम्राजीव्या ना जोन	rapitati anta	. 3	7.comp.clan -agair§1	טשטע	B.com/ny.clair rrustaa	n ngha panh a ngha 199		er ater 1234	T.e	ട ന്നപ്പിന്നം പ്രമന്ത്ര 2000 -	₫IJų
AM 99	. U	234	200	• ∠	9 ⁰ gr	33		1234		2000	
8.வஸ்டிய்கிஅறுபற் எஃப்.4 9 ரம் வரம் ஆண்	പിയ ന്ന്നും ന്ന്	ն առցան	alain (j)		8.ca minyusi m ar Adusi B <u>asti</u>	algung orgh	autori (C	a.dadi Ogal	1 120	കൾ	
17 06 2	010	2 1 2	20	1.1	17	DB	2010	12	92	2011	
aufi G aal orga Aufi	រភាពប ខ្មែរថ្នាំ	19 orgé	3.410		ant Cast	orgo	a#0	ខ ៧៤វិច្ឆា 9 ១១	orgó	a.mg	
23 04 1	012 2.		2 .	a 9	03	04	2016	20	12	2006	
9. s. tudi ciyê ên ejgi cerasî					e fuina.¢	And a Siling of				8	
organi	ಿಕೆ ನೈ ಹಾವ್			1.1	Rget		Guedreaer				



Speech Technology:

Speech Corpus:

Annotated Speech Corpora of approximately 50 hours developed for Hindi, Marathi, Punjabi, Bengali, Assamese, Manipuri, Tamil, Malayalam, Telugu, Kannada.

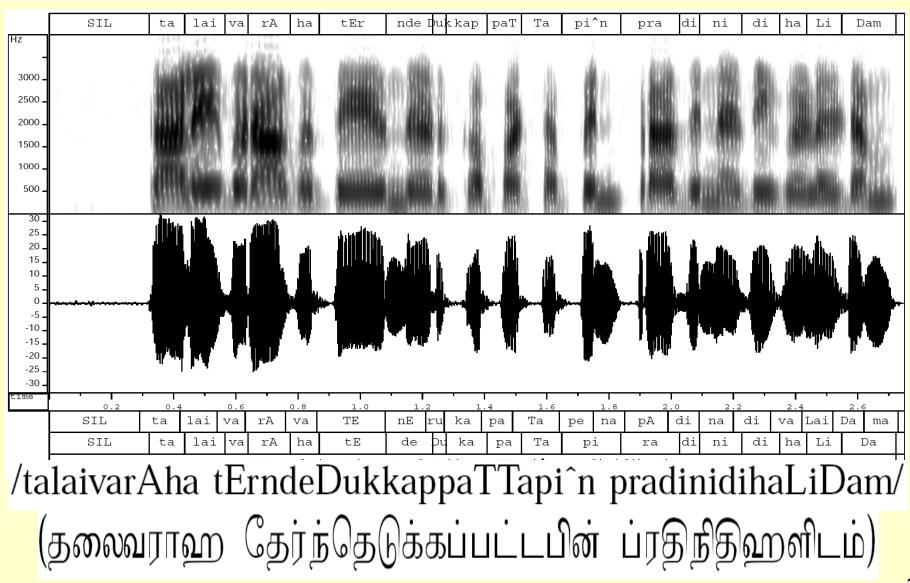
Text –to-Speech for 6 Indian Languages Hindi , Bengali, Marathi , Tamil , Telugu and Malayalam is being developed for visually challenged section of society

Speech Recognition:

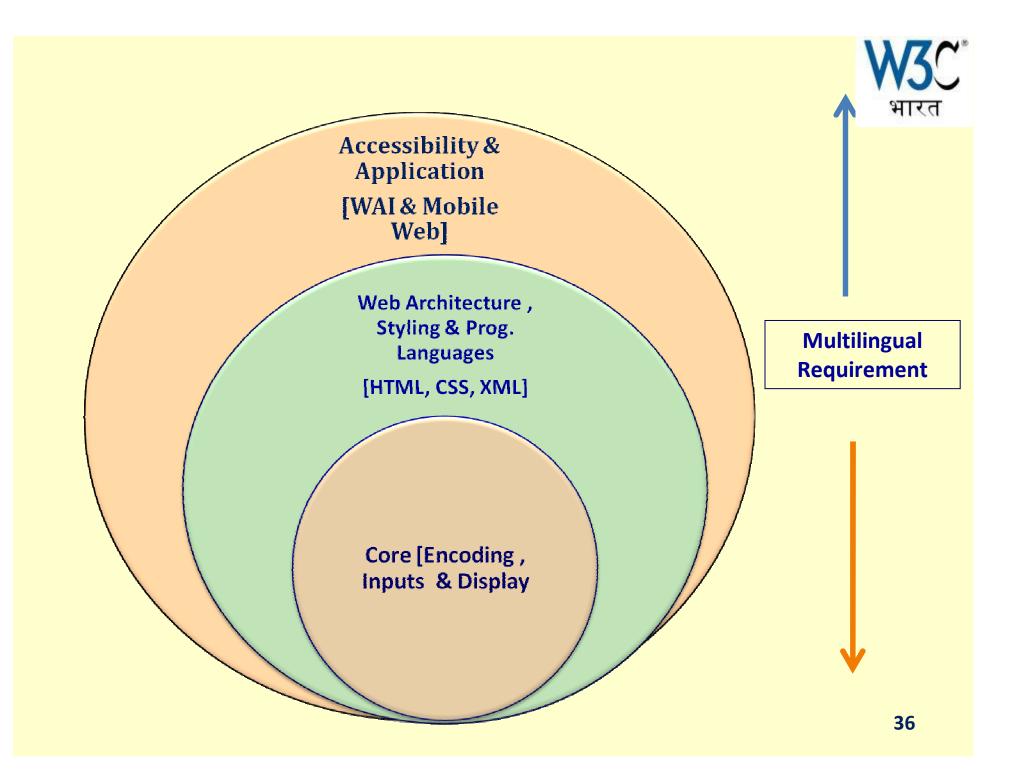
Consortium Mode Project has been initiated for development of Automatic Speech Recognition systems for agricultural commodity prices in six Indian Languages : Hindi, Tamil, Telugu, Bengali, Assamese and Marathi languages

Phonetic Engine for Speech recognition system for Hindi and Telugu languages are being developed.





Standardization Aspects for Supporting Multilingual Web in Indian Languages



W3C^{*} भारत

Character Encoding : UNICODE

- Basis of Multilingual Web.
- All data exchange would be possible seamlessly across devices and Platforms
- Unicode Encoding for all 22 Constitutionally Recognized Indian Languages Complete.
- Unicode declared as a Standard for Data Storage for Web Based E-Governance Services in India

Styling Issues in Indic Languages

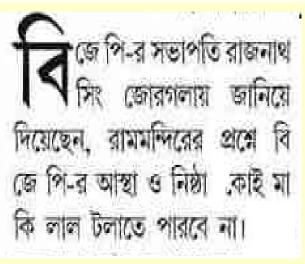


Drop Letters in Indian languages

 Issues for Indian Languages with respect to first character used in Hindi, Malayalam, Bengali, Telugu and Gujarati etc.

केट के लिए मशहूर व इसके लिए दीवानगी दिखाने वाले खेल प्रेमियों के देश भारत में अगर कोई अन्य खिलाडी प्रशंसा हासिल करता है तो यह किसी उपलब्धि से कम नहीं है। फिर हम जिस खिलाडी की बात कर रहे हैं उसने न केवल प्रशंसा अर्जित की है, बल्कि स्टारंडम भी बनाया है। लोग अब उन्हें उनके नाम से जानते हैं और यही अर्जुन अटवाल की क्रिकेट क्रेजी देश भारत में सबसे बडी उपलब्धि है। गोल्फ कोर्स में तो अटवाल अव्वल खिलाडियों में हैं ही। अटवाल ने जब अमेरिकी पीजीए

ഈ ലേഖകനെ വളരെയേറെയാ കർഷിച്ചിട്ടുള്ള വിലാപകാ വ്യമാണു കണ്ണുനീർത്തുള്ളി. നാലപ്പാട്ട് നാരായണമേനോന്റെ ഈ വിലാപകാ വ്യവും വിക്റ്റർ യുഗോവിന്റെ നോവൽ വിവർത്ത നമായ 'പാവങ്ങ'ളും



ચૂંટણી કમિશનર એસ. વાય. કુરેશીએ ભારતમાં દ્વો કસભાની ચૂંટણીની સંભવિત તારીઓનો લંડનમાં નિર્દેશ આપ્યો અને ભારતના રાજકારણીઓમાં હલચલ મચી ગઈ. વાસ્તવમાં એ ચૂંટણીપંચની સત્તાવાર જાહેરાત નહોતી, પરંતુ હવે રાજકીય પક્ષે એવું માનવા લાગ્યા છે કે ચૂંટણીની તારીઓની ગમે ત્યારે જાહેરાત થઈ શકે છે. આવું થાય તો એ દિવસથી ચૂંટણીની આચારસંહિતાનો અમલ શરૂ થઈ જાય. શાસક મોરચાના પક્ષોને અને ખાસ કરીને સરકારનાં વિવિધ મંત્રાલયોના મંત્રીઓએ ચૂંટણી પહેલાં લોકોના લાભની અનેક યોજનાઓની જાહેરાત કરવાની તૈયારી કરી રાખી છે. మవారం మార్కెట్ భారీగా లాభపడింది. రావడంతో సెన్సెక్స్ పైఫైఫై పోయింది. సెన్సెక్స్ -283.03 పాయింట్ల లాభపడి 9583.89 వద్ద ముగిసింది. నిష్టి 76.80 పాయింట్ల లాభపడి 2910.90 పాయింట్ల వద్ద ముగిసింది. లాభపడిన రంపెనీల పేర్న ఫెల్, హెచ్డిఎఫ్స్, హెచ్డిఎఫ్స్బ్యాంక్, ఇన్ఫోస్, ఎల్ఎన్టి, మారుతీ, రిలయన్స్, ఎస్టిబ, టాటామోటార్స్, టాటా పవర్స్, టిసిఎస్, విణిబ, టాటామోటార్స్, టాటా పవర్స్, టిసిఎస్, విణి, బిబిస్బ ఖ్యాంక్, ఐటిసి, ఎన్టిపిసి, ఒఎన్టిసి, వంటే రంపెనీలు లాభపడాయి

Underlining of characters



 There is some examples of Indian languages in which Matra's are not readable due to underlining of characters

Hindi -	अन्य भाषाओं में भी अनुवाद				
Punjabi-	ਗੁਰੂ				
Bengali-	<u>ভাই পুরোনো আর্কাইভ একটু ওলট পালট।</u>				
Guajarati - <u>सरदार ગુર્જરી</u>					
Marathi-	<u>मराठी मुला मुलींची नावे</u>				
Tamil-	நீரிற்குமிழி யிளமை நிறைசெல்வம்				
Telugu -	<u>శ్రీశైలం ప్రాజెక్ట్రపై TV9 ప్రోగ్రాం ''' డ్యామ్ ఇస్ డేంజర్ '' పార్ట్</u>				



• Vertical arrangements	चाँ द		व क्ता	or	व क् ता	or	वक् ता		श क्ति	or	श क् ति
• Bullets and Numbering		अ) इ) ई) उ) उ) ए) ऐ)		+09 +09 +09 +09 +09 +09 +09 +09	06 07 08 09 0A 0F 10			ਆ (-0A73 -0A05 0A72 0A38 0A39 0A15 0A16 0A16 0A17 -0A18 0A19	
		औ)		+09	14			ड)) U+ 	0A1B -	

Indentation of character



Under Lining of the characters



C: Documents and Settings\temp851\Desktop\blink.html - Microsoft Internet Explorer	
File Edit View Favorites Tools Help	1
🔇 Back - 🕥 - 🖹 🗟 🏠 🔎 Search 👷 Favorites 🚱 🍰 👼 - 🛄 🚛 🏭 🍇	
Address 🖗 C (Documents and Settings)temp651 (Desktop)blink.html 👻	🛃 Go Links 🎽 🍖 Convert 🔹 🕵 Select
Hindi :	<u>^</u>
सङ्गंडन खेलों के लिए स्टेंडिका सौंबने को तैयार नहीं है	
English :	
Commonwealth Games stadium is not ready for handover	
Bangla :	

Tamil :	
பாப்லி அணை, விவகாரம் , தொடர்பாக	
Gujrat :	
લેંદરાય ઝામાક ત્રેપુલિણ-ગ્રેલર-વાળા તેતાય તેલેવ ગાહ્ય વંદુવા ગાણ્યુ	
Marathi :	
बांकळी थरणावर तेतुनु देरागचे अध्वया चंद्रावानू नावडू वांचा स्तलेला राजकीय-काठा	
Punjabi :	
สมมีลสมับสอบไลต์ออีกรายและเวลาเราราน	
Malyalam	
800 0.8. 3800.6.3805.0.803;805	
Kannada:	
addaddaadaddadaaad, foraaad, j	
	V
Done	💡 My Computer
🛃 start 🔰 🥸 🎒 🁋 📎 Ind 276, 5 🔮 Moella Fir 👔 blink-Not 🔞 2 Maros 🏷 Marosoft 🤌 2 Intern 🥞 unbib	led 🛛 HI 😰 🗘 🌾 🏯 🐠 🗭 2:43 PM

Major Identified Problems in Styling :

- Grapheme Cluster Problems for Vertical Writing Style
- Drop Initial Views of the First Letter Element
- Bullets & Numbering issues
- Justification Problems
- Horizontal Letter Spacing Problems
- Most browsers are unaware of syllable boundaries for Indic scripts.
- Current workaround for this is enclosing the first syllable in a separate span governed by a dedicated CSS class.
- Mozilla Firefox seems to recognize it well.



Approach to be taken for Possible Solution



1.Grapheme Cluster Problems :

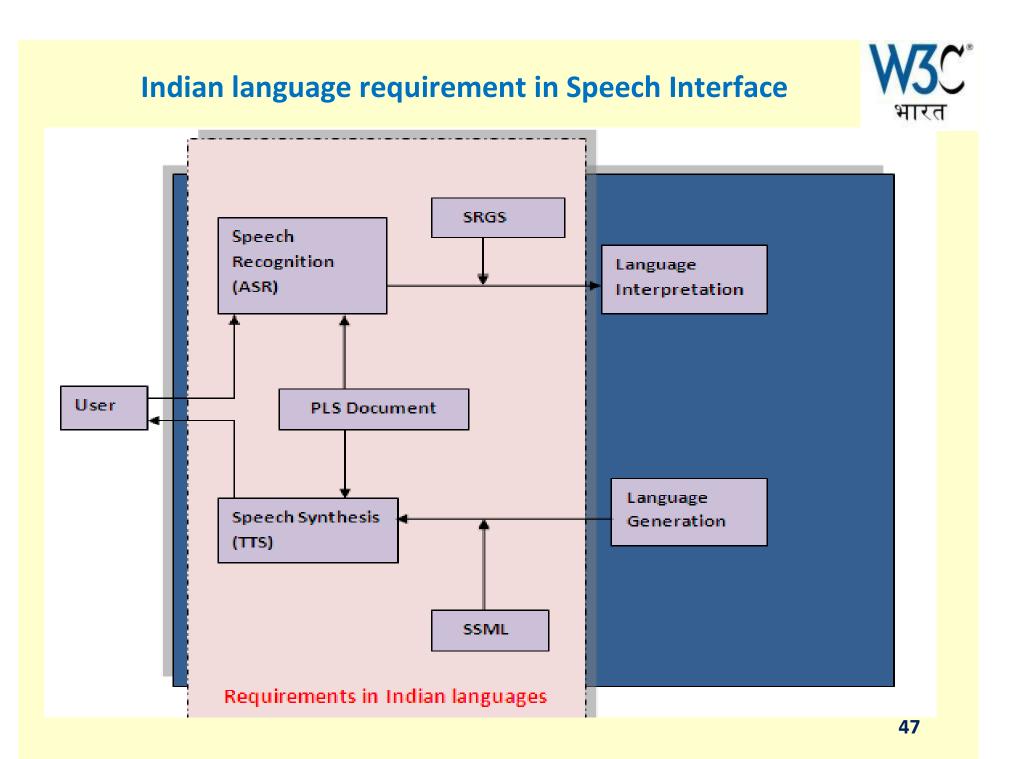
Adoption of UAX#29 (Unicode Text Segmentation Algorithm addressing the complexities of Indian Grapheme Clusters)

- 2.Bullets & Numbering issues
- **3.Justification Problems**
- 4.Horizontal Letter Spacing Problems
- Development of Complete Mapping Table involving detailed requirements for document layout, typography and typesetting and calligraphic conventions, i.e. Styling Manual.
- The Rules developed in Styling Manual needs to be converted for inclusion in HTML and CSS



- An activity involving in-depth analysis of browser rendering engine needs to be taken up on a urgent basis.
- A reference web-page layout engine needs to be developed which could act as a benchmark for other browser players to emulate.
- Mozilla Firefox could be taken up for modification to act as a reference implementation.
- Core Working Group under the aegis of W3C Internationalization Group being formed.

Enabling speech interface for web



W3C Pronunciation Lexicon Specification (PLS 1.0) – Voice Browsing



W3C Voice Browser Activity, has published a Pronunciation Lexicon Specification (PLS) Version 1.0. PLS is designed to enable interoperable specification of pronunciation information for both speech recognition and speech synthesis engines within voice browsing applications.

PLS Attributes:

Multiple pronunciations

- For ASR: to accommodate speaker/regional variability, not native speakers
- For TTS: a single preferred pronunciation will be selected
- Multiple orthographies (with same pronunciations)
- Useful for both ASR & TTS
- Homophones (same pronunciations, different meanings)
- Homographs (same spellings, different pronunciations)

Proposed Approach for incorporating Multilingual requirement in PLS:

Usage of Parts of Speech (POS) information for resolving multiple pronunciations

- Parts of Speech (POS) plays an important role in Indian languages (like Bangla, Hindi) pronunciation.
- Based on the POS same orthography can produce different pronunciation.

Pos as an attribute:

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"</pre>
   alphabet="ipa" xml:lang="bn">
 <lexeme>
  <grapheme>সরল</grapheme>
  cphoneme pos= "adjective"> /pro/
  <!-- IPA string is: " forof" -->
  <!---Meaning is : "easy" -- >
  <phoneme pos= "verb" > forlo</phoneme>
  <!-- IPA string is: " forld" -->
  <!---Meaning is: "moved" -- >
  <phoneme pos= "null"> forol</phoneme>
  <!-- IPA string is: " forof" -->
  <!---Itrans is: "sarala" -->
  <!---Meaning is: "easy" -->
 </lexeme>
</lexicon>
```

Usage of morphological information for resolving multiple pronunciations

- In Indian Languages not only POS information but also morphological information are very crucial in determining the pronunciation of a homograph.
- This information can be defined in the same attribute or element used for POS using proper POS tag set for that language.

Example : Bengali

Web Accessibility & Implementation



Challenges :

- Make Web content accessible to people with disabilities w.r.t Indian languages
- WCAG 2.0 Guidelines for success criteria vis-a-vis selected recommendations relevant to Indian context

Initiatives in India :

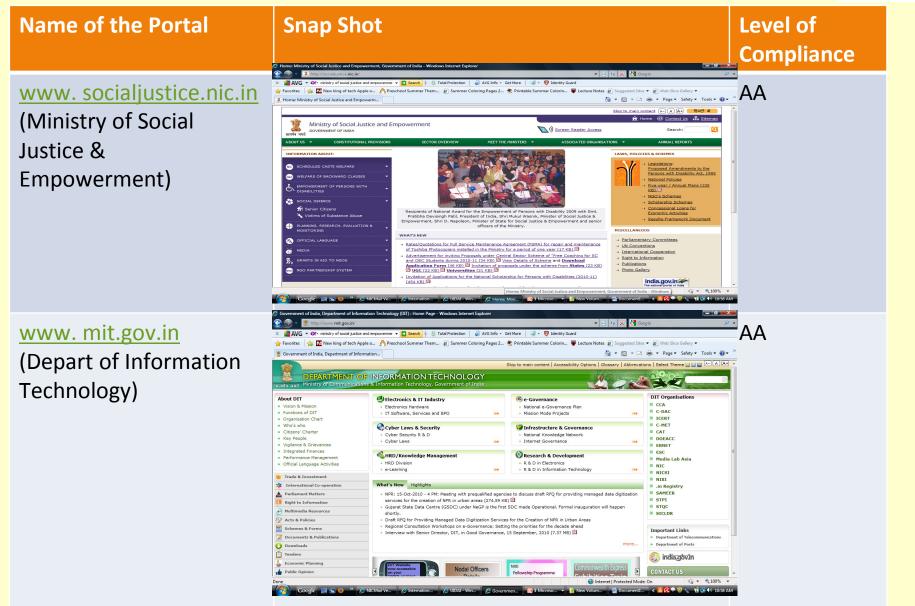
- "Guidelines for Indian Government websites" by NIC , Govt. of India
- STQC Implementing WCAG 2.0 Accessibility through Website Quality Certification
- Centre for Internet and Society developing authorized translation of WCAG 2.0 Guidelines

Roadmap :

- Meet WCAG 2.0 guidelines & techniques w.r.t Indian languages
- Initiation with Hindi, Bangla, Marathi, Telugu, Tamil

Some of the WCAG 2.0 Compliant Web Sites







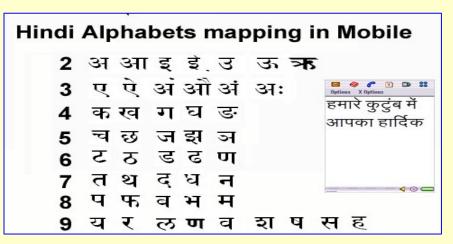
Issues for enabling Mobile Web in Indian languages

- Character encoding
- Bandwidth and Cost
- Backward Compatibility with Legacy Devices
- Lack of standardization
- Fonts
 - Bit map fonts (used by low cost handset)
 - True type fonts (used by high end handsets)
 - Open type fonts (currently in wider use)
- Common Storage format
- Mobile messaging in indic languages
- Presentation Issues

Mobile keypads issues

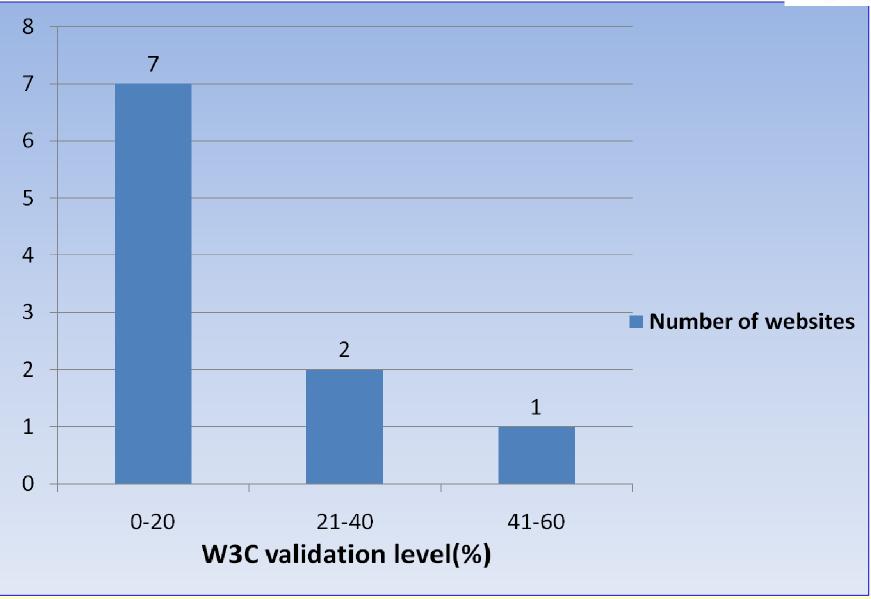


- Multi-tap issues
- Too many taps per key for each char No way to know which char is on which Key.
- Never support more than one language on the keypad because there is not enough space on the key face to print more characters.



- Dictionary Based
- Transliteration

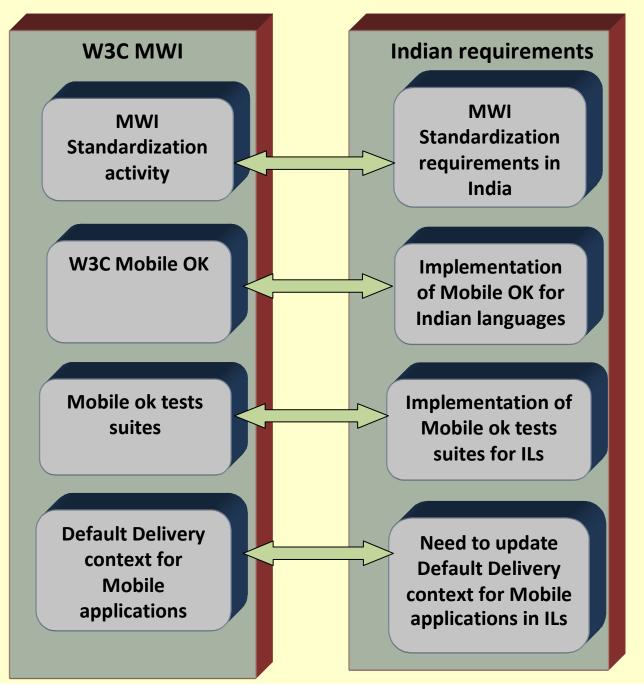
Level of Adherence to W3C Mobile Ok while testing 10 Hindi websites

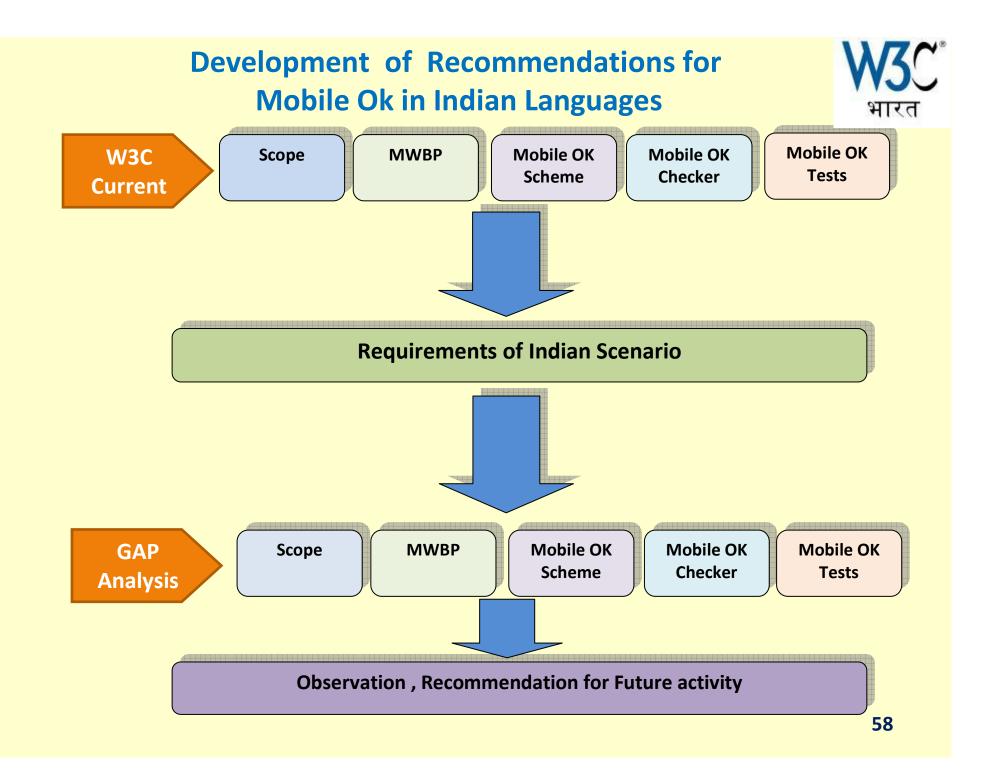


भारत

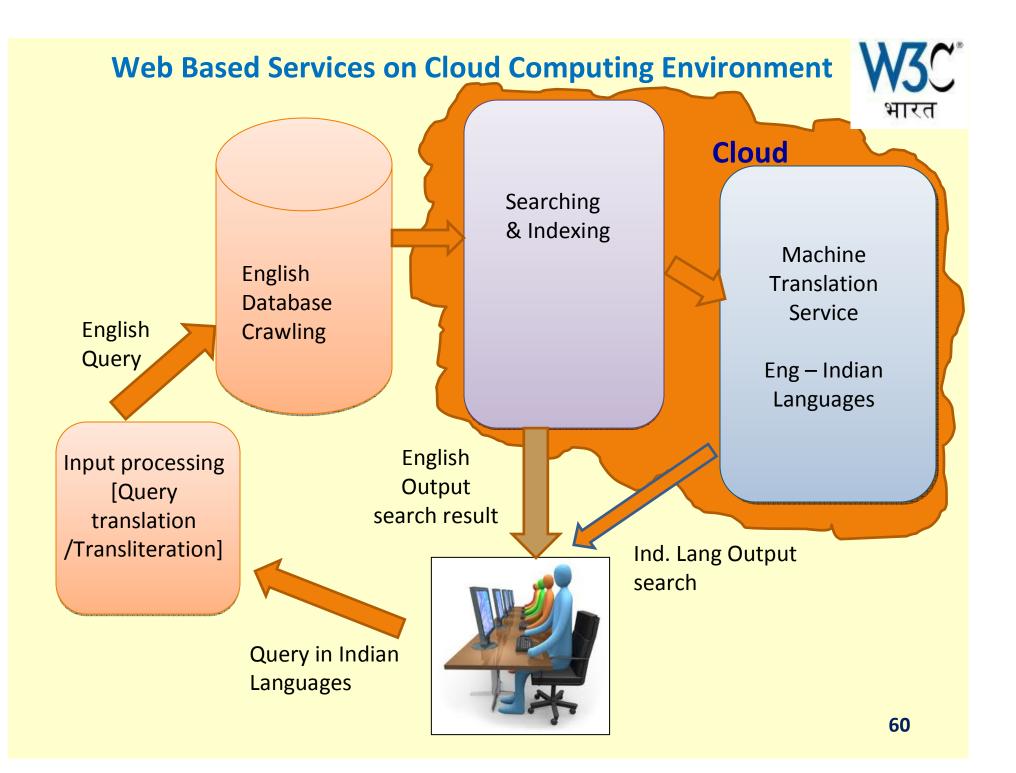
Proposed approach: Gap Analysis







Future Initiatives

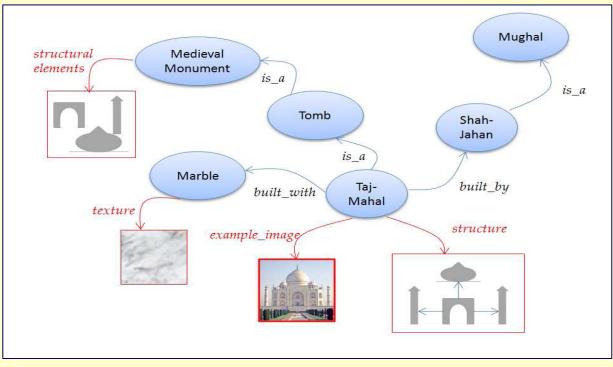


Semantic Web

1. Development of Word-Net for Indian Languages.

W3C^{*} भारत

- 2. Development of Multimedia Web Ontology Language (MWOL)
- Multimedia Web Ontology Language" (MOWL), is designed as an extension of OWL, the W3C recommended ontology language for the web.
- MOWL supports creation of and reasoning with perceptual modeling of concepts, and probabilistic evidential reasoning.



Formation of the concept Indian "Medieval *Monument*" and the abstracted visual patterns that are expected on an of embodiment the concept in a multimedia artifact.

Some of Future Initiatives:



- Multilingual Requirements for Indian Languages for HTML 5.0
- Multilingual requirements for Voice Browser in Indian Languages
- Gap Analysis for implementation of Mobile Web best practices in Indian Languages
- Multilingual Linked Data for E-Governance data

The Internationalization Best Practices Would be foundation of the all the above initiatives

Thanks & Questions

