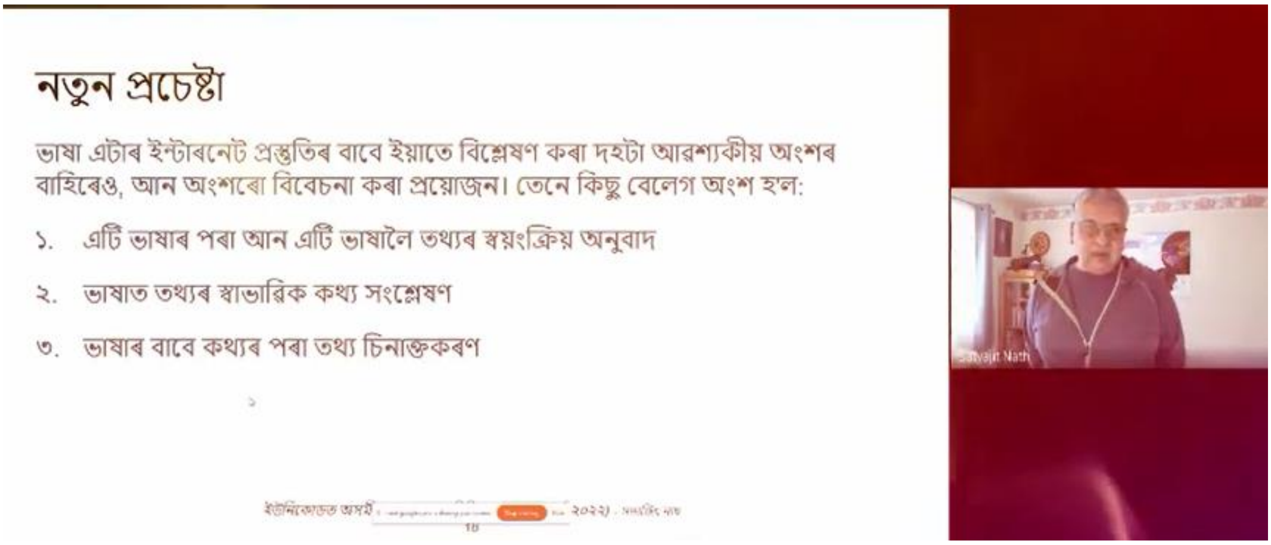
Workshop held on
“Assamese Unicode and Various Challenges “

One technical workshop viz. Assamese Unicode and Various Challenges by Sri Satyajit Nath, an experienced software professional from USA, was held on 13th March 2022 from 7:30 PM IST onwards through online mode where around 100 nos of interested persons participated from various parts of the world. The workshop was organised by Assam Association Delhi.

During the hour long presentation in Assamese, Sri Nath spoke about various topics viz., how computers and mobile phones use codes to store various letters of scripts, What is Unicode, How it is different from earlier codes viz. **ASCII ,** ISCII , ISO-8859-1, Unicode for Assamese *language*, Assamese keyboards and fonts, and the various challenges it is facing viz, Auto Translation, transliteration, language-specific processing viz. Synonyms and Antonyms suggestions, auto spelling correction, search engine optimization for Assamese language in Google, Bing and how to enhance capabilities for future requirements viz. AI, Machine Learning, Robotics, Data Analysis, Big data Analytics etc.



নতুন প্ৰচেষ্টা

ভাষা এটাৰ ইন্টাৰনেট প্ৰস্তুতিৰ বাবে ইয়াতে বিশ্লেষণ কৰা দহটা আৱশ্যকীয় অংশৰ বাহিৰেও, আন অংশৰো বিবেচনা কৰা প্ৰয়োজন। তেনে কিছু বেলেগ অংশ হ'ল:

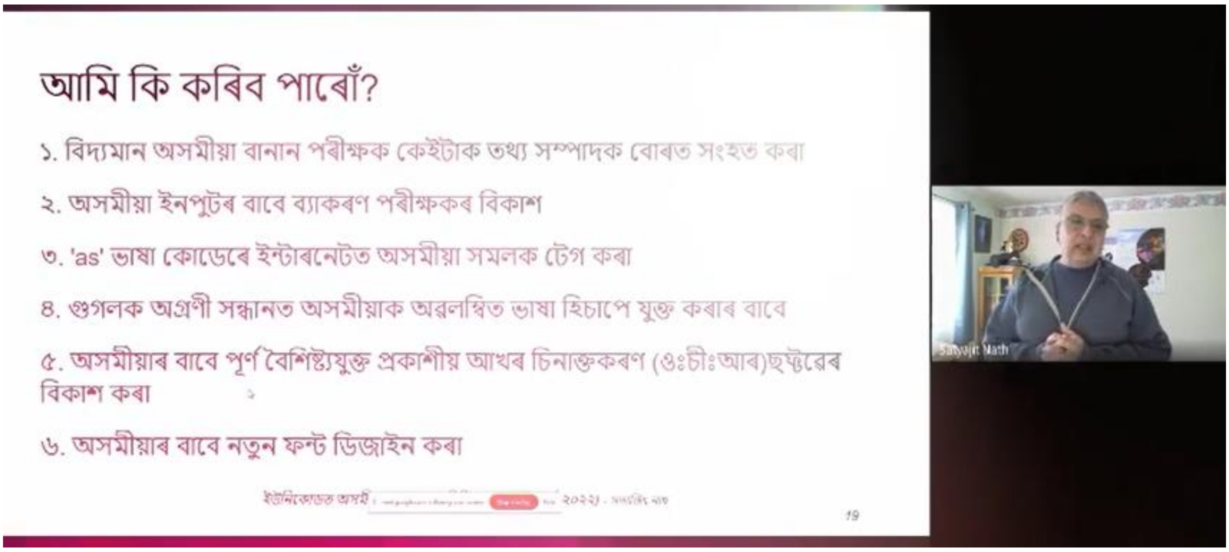
১. এটি ভাষাৰ পৰা আন এটি ভাষালৈ তথ্যৰ স্বয়ংক্ৰিয় অনুবাদ

২. ভাষাত তথ্যৰ স্বাভাৱিক কথ্য সংশ্লেষণ

৩. ভাষাৰ বাবে কথ্যৰ পৰা তথ্য চিনাক্তকৰণ

ইউনিকোডত অসমী ২০২২) - সত্যজিৎ নাথ

18



আমি কি কৰিব পাৰোঁ?

১. বিদ্যমান অসমীয়া বানান পৰীক্ষক কেইটাক তথ্য সম্পাদক বোৰত সংহত কৰা

২. অসমীয়া ইনপুটৰ বাবে ব্যাকৰণ পৰীক্ষকৰ বিকাশ

৩. 'as' ভাষা কোডেৰে ইন্টাৰনেটত অসমীয়া সমলক টেগ কৰা

৪. গুগলক অগ্ৰণী সন্ধানত অসমীয়াক অৱলম্বিত ভাষা হিচাপে যুক্ত কৰাৰ বাবে

৫. অসমীয়াৰ বাবে পূৰ্ণ বৈশিষ্ট্যযুক্ত প্ৰকাশীয় আখৰ চিনাক্তকৰণ (ওঃচীঃআৰ)ছফ্টৱেৰ বিকাশ কৰা

৬. অসমীয়াৰ বাবে নতুন ফন্ট ডিজাইন কৰা

ইউনিকোডত অসমী ২০২২) - সত্যজিৎ নাথ

19

During 'Question Answer' session, Sri Abhijit Dutta, 6th Semester Degree student, Assamese Department of North Lakhimpur College, Assam enquired why Geetanjali font has stopped working properly in MS-Office? .

In response, Sri Nath informed Geetanjali font is not an Unicode based font and currently modern applications don't support non-unicode based fonts including Geetanjali font in their application including MS Office. Documents prepared with Geetanjali fonts use ISO-8859-1 codes meant for European letters instead of Unicode codes for Assamese letters. This prevents the ability to search or do processing of the text in those documents. So, Geetanjali font should not be utilized in the internet era now. But if there are some content which have already been developed using Geetanjali font, these should be converted into Unicode using some conversion tools; some of which are available freely in internet . Nowadays software companies like Microsoft have developed Unicode keyboard for Windows, viz. Windows Assamese Inscript keyboard, which can be used with Microsoft word, excel etc. for writing in Assamese. Microsoft includes Unicode compliant fonts that support Assamese in word, excel, etc.

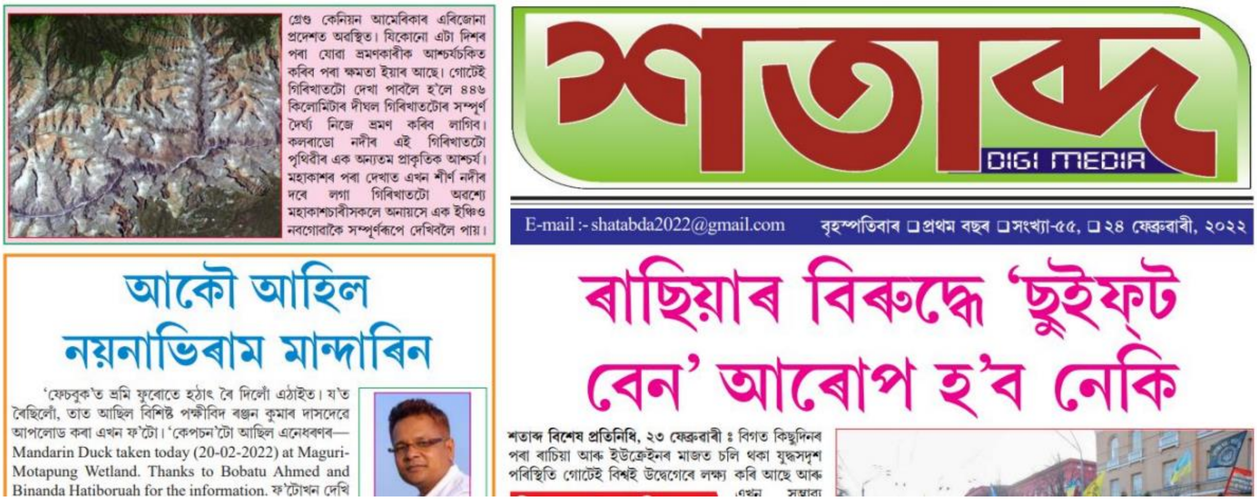
Smti Syeda Jebeen Shah , New York, USA informed that as large numbers of Assamese words or sentences are being wrongly fed into Google and Microsoft platform by citizen, so several Assamese words or sentences are wrongly displayed by Google, Microsoft during language processing viz. searching, translation, synonyms, antonymous etc. As these platforms yield results on the basis of inputs received through crowd-sourcing, so it was requested by her to contribute only correct and verified data. She also informed that she has so far contributed 10,016 nos Assamese words in Google platform and still contributing and encourage other to contribute as much as they can.

In response, Sri Nath conveyed that though big companies viz. Google, Microsoft have been involving in language translation project including Assamese but their product quality depend heavily on data contributed by community. They generate output on the basis of input data received from public after processing through Artificial Intelligence, Machine learning, etc. So, it is important to feed correct input data to generate correct output, otherwise it will be GIGO – Garbage in , Garbage out. So we should encourage to feed correct data only. But one significant issue is that these language processing initiatives have been so far done by private companies viz. Microsoft, Google which are mandated by their own policy. They may or may not follow as per requirement of citizen. So it is important to involve Non Profit Organization using Open Source Technology to ensure more transparency and accuracy. As IIT Mumbai has been working for Hindi, similar approach may be taken up for Assamese involving premier Institutions of Assam.

Then Query was raised by Dibyojit Dutta how to retrieve so much of content prepared using non-Unicode codes particularly media house where large nos of newspapers, magazines are still using non- standard codes which are inaccessible through search engine in internet platform.

In response, Sri Nath conveyed that there are still several Assamese newspapers /magazine which are still using DTP SW with non-unicode fonts and then publish after converting into pdf or image file. these are not been able to be searched in internet and used for other purposes. These newspaper publishers should immediately start using Unicode text and fonts as a few newspaper have been seen to be shifted successfully and their success stories can be utilised. For old and existing content which have already been published as PDF using non-compliant fonts or image copy, these content may be extracted using OCR SW

গীতাঞ্জলি ফন্টৰ বিষয়ে



গ্ৰেণ্ড কেনিয়ন আমেৰিকাৰ এৰিজোনা প্ৰদেশত অৱস্থিত। যিকোনো এটা দিশৰ পৰা যোৱা ভ্ৰমণকাৰীক আশ্চৰ্যচকিত কৰিব পৰা ক্ষমতা ইয়াৰ আছে। গোটেই গিৰিখাতটো দেখা পাবলৈ হ'লে ৪৪৬ কিলোমিটাৰ দীঘল গিৰিখাতটোৰ সম্পূৰ্ণ দৈৰ্ঘ্য নিজে ভ্ৰমণ কৰিব লাগিব। কলৰাডো নদীৰ এই গিৰিখাতটো পৃথিৱীৰ এক অন্যতম প্ৰাকৃতিক আশ্চৰ্য। মহাকাশৰ পৰা দেখাত এখন শীৰ্ণ নদীৰ দৰে লগা গিৰিখাতটো অৱশ্যে মহাকাশচাৰীসকলে অনায়াসে এক ইঞ্চিও নবগোৱাকৈ সম্পূৰ্ণৰূপে দেখিবলৈ পায়।

শতাব্দ

DIGI MEDIA

E-mail :- shatabda2022@gmail.com বৃহস্পতিবাৰ ❑প্ৰথম বছৰ ❑সংখ্যা-৫৫, ❑২৪ ফেব্ৰুৱাৰী, ২০২২

আকৌ আহিল নয়নাভিৰাম মান্দাৰিন

'ফেচবুক'ত ভ্ৰমি ফুৰোতে হঠাৎ বৈ দিলোঁ এঠাইত। য'ত বৈছিলোঁ, তাত আছিল বিশিষ্ট পক্ষীবিদ ৰঞ্জন কুমাৰ দাসদেৱে আপলোড কৰা এখন ফ'টো। 'কেপচন'টো আছিল এনেধৰণৰ— Mandarin Duck taken today (20-02-2022) at Maguri-Motapung Wetland. Thanks to Bobatu Ahmed and Binanda Hatiboruah for the information. ফ'টোখন দেখি

ৰাছিয়াৰ বিৰুদ্ধে 'ছুইফ্‌ট বেন' আৰোপ হ'ব নেকি

শতাব্দ বিশেষ প্ৰতিনিধি, ২৩ ফেব্ৰুৱাৰী ঃ বিগত কিছুদিনৰ পৰা ৰাচিয়া আৰু ইউক্ৰেইনৰ মাজত চলি থকা যুদ্ধসদৃশ পৰিস্থিতি গোটেই বিশ্বই উদ্বেগেৰে লক্ষ্য কৰি আছে আৰু এখন সম্ভাব্য

As several Assamese newspapers are still preparing their content using Non- Unicode DTP SW, converting into pdf or image file and publish , these content can neither be internet-searched not utilised in other language processing services. These newspapers should switch over to Unicode. As a few newspapers have already shifted successfully to Unicode , their success stories should be followed. For content which have already been published as PDF or Image , these words may be extracted using OCR SW. Also the original Softcopy available in non- standard code may be converted into Unicode using some Conversion tools – some are Open Source freely available and some are proprietary . Ramdhenu Company has already developed tools to convert Gitanjali font into Unicode. Sentinel Group of Newspaper is one example of a newspaper company that has been publishing their Assamese online newspaper ‘`Assamese Sentinel` https://assamese.sentinelassam.com/ in Unicode. So other newspapers may learn from their experiences.

On query , on various initiative taken on Assamese language processing by various agencies , it was told that several activities have been seen in various platform to support Assamese language processing viz. fonts, keyboard, synonymous/ antonymous database , language translation, etc. Xobdo.org has been developing database on NER language dictionary . Similarly a few are working in unicode conversion, font, keyboard etc viz. Jhanavi.net, Society for Technology Development of Assam etc.

শব্দ Xobdo.org	Jahnabi Multilingual Input Tool 2.0.5 জাহ্নৱী ২ বহুভাষিক লিখন সঁজুলি Build your own keyboard layout for any language supported by Unicode/non-Unicode or use the bundled ones. Jahnabi supports Inscript, Phonetic etc. for details visit: www.jahnabi.org www.facebook.com/groups/jahnabi.soft © Utpal Phukan d_Ant <defunq@outlook.com 1/5	ভাষা প্ৰযুক্তি বিকাশ পৰিষদ, অসম Society for Language Technology Development, Assam প্ৰযুক্তিৰ শলিতাৰে জ্বলাওঁ ভাষাৰ চাকি

Now it is important that the products /contributions of various initiatives should be linked with products of Microsoft, Google, Apple, Adobe etc for wider usability. As these big companies have been publishing API (Application Programming Interface) to integrate 3rd party functionality, such integration through API's to be made by our agencies.

Smti Bhiba Bharali from Assamese dept, Gauhati University has informed that though PHD students of Assamese department of GU are presently using Unicode, but previously Ramdhenu SW was used for their research papers. But when direction was issued to students to resubmit their research papers after converting into Unicode probably to check or prevent plagiarism, lots of students faced difficulty in conversion i.e. conjunction not properly rendered, broken form, junk characters etc. So enquired any speedy solution on this issue

In response, Sri Nath conveyed that the difficulty in conversion might be due to faulty software tools used. Alternatively they can convert their old text to Unicode and then edit using a Unicode compliant font to correctly display the text in Assamese. Any illegible portion remaining can be edited using a Unicode compliant keyboard such as Windows Assamese Inscript which is provided by Microsoft on Windows. The Ramdhenu company which is using Geetanjali font has developed one SW tool to convert text using Geetanjali font into Unicode text. They have also developed one keyboard for Assamese with layout similar to their previous keyboard to generates Unicode codes. So those students who are conversant with Ramdhenu software may change their keyboard to this one.

Sankar Krishna Das enquired provision to preserve ancient books through scanning. Sri Nath conveyed that Asom Sahitya Sabha has already working for preservation / conservation of ancient books/ manuscripts through scanning involving AMTRON , but extracting of the content to machine readable words or sentence is much required now. This would be possible through using OCR SW to convert those scripts into Unicode. And in order to recognise a few ancient scripts, some modification may be required in the SW to recognise correct output. Some of our start-ups can look into it to develop such SW. Some reputed SW companies who have been working on OCR SW may be approached but they may see the commercial benefits.

Sri Sankar Krisha also suggested the requirement of fonts similar to sachipath scripts which should have commercial value to utilise in special purposes viz. marriage invitation letter etc .

While discussing , if it is possible to extract meaning from manuscript through OCR Scanning, it was felt that the existing OCR SW may not be able to recognise as the manuscript script were different at that time for which it may require special OCR. Sri Pranab Phukan , Delhi expressed that inspite of recognising of script and conversion into text from sachipath or ancient books, the existing language processing SW may still find it difficult to generate proper meaning as words, sentence had different meaning at that time. So, it would require advanced analysis.

At the end, Sri Dibyojit Dutta who moderated the event conveyed that when Land records Computerisation project was implemented in the country across all States in Nineties, then 7 bits codes were utilised to store Assamese scripts including other Indian Scripts. But when Unicode came around 2000, those data in 7 bits codes were converted into Unicode through conversion tools. If 90-95% data can be retrieved, the remaining 5-10% can be corrected through text editors. So there is no reason , why Assamese couldn't convert their large pool of content into Unicode, and make these available in internet as searchable for the larger benefit of the people.

During the Vote of thanks , Sri Dutta expressed their gratitude to Sri Satyajit Nath who had provided valuable insights into various aspects of Assamese Languages in Unicode and their challenges. He further expressed that the knowledge and expertise of Assamese Diaspora can be leveraged by various institution in implementing language specific projects.

The entire session was recorded and made available in the link https://www.youtube.com/watch?v=NWZtcljAt9k

Dibyojit Dutta
GS, AAD
2nd April 2022

Attachment

a. Power point presentation
b. Resource materials on Assamese in Unicode