# Challenges in building multilingual multidirectional lexical search - the case of Nyishi-Bangla-English trilingual lexicon

Atanu Saha

Prof. Girish Nath Jha

Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi

# Introduction

- Nyishi, a language spoken in Arunachal Pradesh and parts of Assam.
- It belongs to Sino Tibetan Family of the Tibeto Burman group.
- Nyishi also has many alternative names such as Nisi, Dafla, Nishing.
- The current population of Nyishi speakers is 118,111.

# Introduction

- The language does not have a written script and currently people believe that this language is endangered.
- To prepare an online dictionary using the data.
- Serving the purpose of giving something back to the community.

# The reasons

- A virtual representation would be helpful for the morphologists, syntacticians, typologists and anthropologists worldwide.
- The speakers of this language can easily see that their language is described and documented.
- Thus they can also further participate in adding more things to it.

## The trilingual data and the challenges in collection and reorganizing the data for computing needs

- Primarily making a dictionary over the web is different from all those conventional printed dictionaries.
- The problem became even severe because Nyishi does not have a script of its own.
- This language is not structurally similar to that of Hindi or English.

# The trilingual data and the challenges in collection and reorganizing the data for computing needs

- For representation of Nyishi data we used the IPA convention.
- Three sets of data
- One containing Nyishi words and sentences in IPA.
- Second with Bangla translation in Bangla Script.
- Third with a translation in English using Roman Script.

# Computer Modules

- The dictionary is made by the java programming platform.
- A typical java system requires an environment, the language, the java application programming interface and various class constructions.

# Computer Modules

- The programme is then saved in the hard disk with .java extension.

- Two major modules of the program - the search engine and the lexicon reader.

- The reader reads a pre specified lexical data as well as the configuration data.

# Programming environment

- The lexical data allows multicolumn search and the configuration data allows limited customization of the system, display etc.

- There are two sets of files containing the data and a configuration file stating the format in which the data will appear.

# Programming environment

- The java server page (abbreviation is .jsp) is a page where the codes of that programming language are written.

- In java programming environment we need to construct certain classes which will be able call the object and return the desired result.

# Programming environment

- Primarily  an excel sheet was used to type the data.

| Number_id | NYISHI [ROMAN SCRIPT | IPA TRANSCRIPTION | Bengali Meaning | English Meaning | CATEGORY | EXAMPLE IN SENTENCE IF ANY |
|---|---|---|---|---|---|---|
| 5 | acin | acin | খাবার | Rice | Noun | ŋa balo acin kayleu gunam diutayin<br>i will eat half of the rice. |

# Programming environment

- After that we had to create two .txt files.
- One is named as config.txt and another is the lexicon.txt in which we kept the same order of the data as in Excel sheet.

# Programming environment

- In the next phase necessary commands needed to be given in order to compile the programme.

- This compiler's (javac) job is to translate the java prog. into byte codes.

- Byte code is the language understood by the java interpreter.

# Programming environment

- Before execution a programme has to be placed in the memory and its done by a class loader.
- The class loader takes the .class files containing byte codes and transfers them into memory.

# Programming environment

- Classes

- Lexicon Reader
- Lexicon Search

# Programming environment

- Before the byte codes are executed by the java interpreter they are verified by the byte code verifier.
- This ensures that the byte codes are valid and they don’t violate java’s security questions.

# The way this dictionary works

- Given a search term at the level of user interface the java environment will call the java classes and if the search is true then the required information will be showed or given as output to the user.
- If the input is null then the result would be no search item is found.
- The programme would not crash because we already have specified the value of null as an exception.

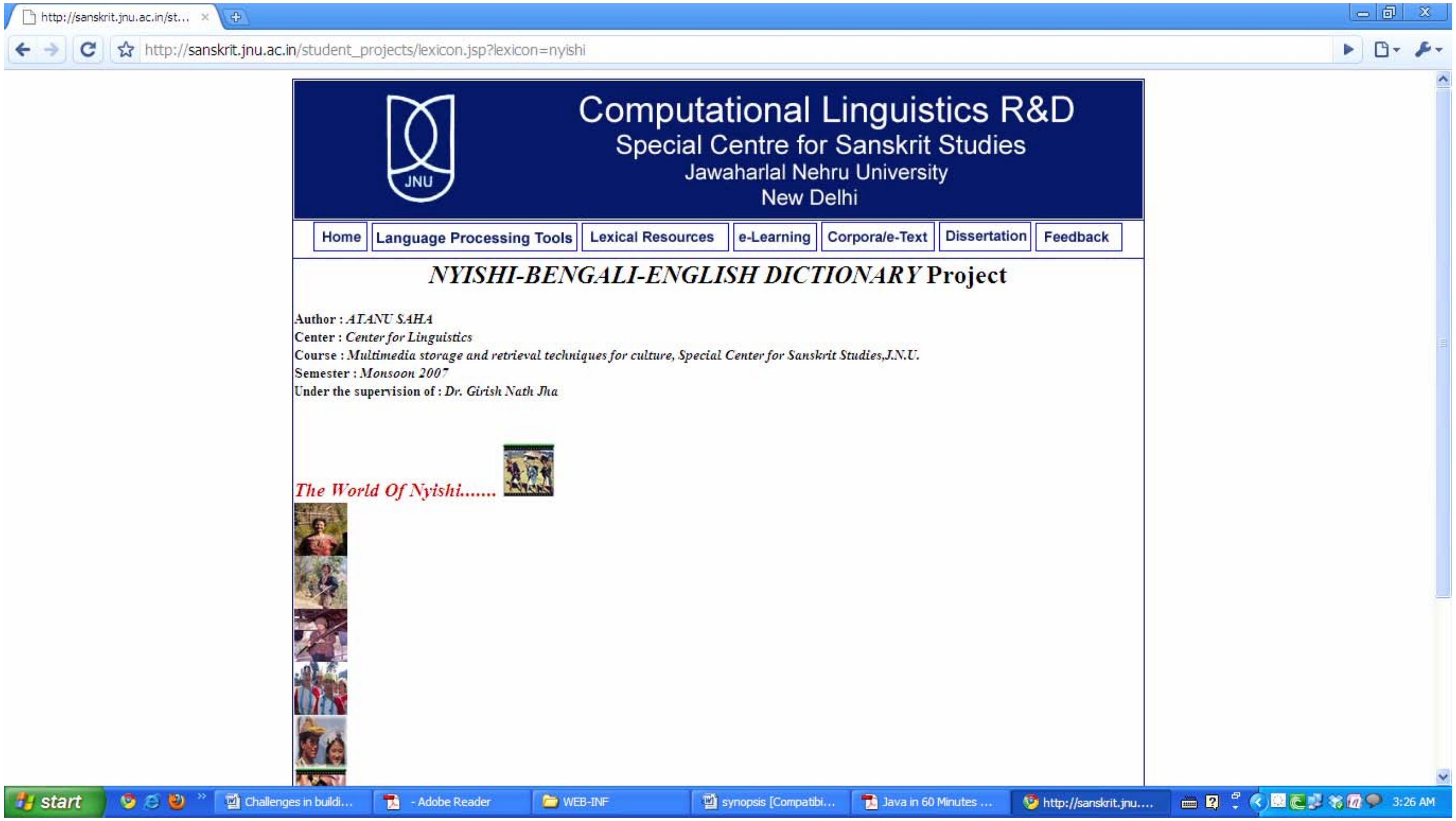
# The way this dictionary works

- A lot of multimedia contents such as images, sound files are also provided to make the dictionary more attractive as well as informative.

- The project is currently available online in **http://sanskrit.jnu.ac.in/course-projects-winter08/nyishi-search.jsp**

# DEMO

# Screen shot 1



http://sanskrit.jnu.ac.in/st...
http://sanskrit.jnu.ac.in/student_projects/lexicon.jsp?lexicon=nyishi
JNU

Computational Linguistics R&D
Special Centre for Sanskrit Studies
Jawaharlal Nehru University
New Delhi

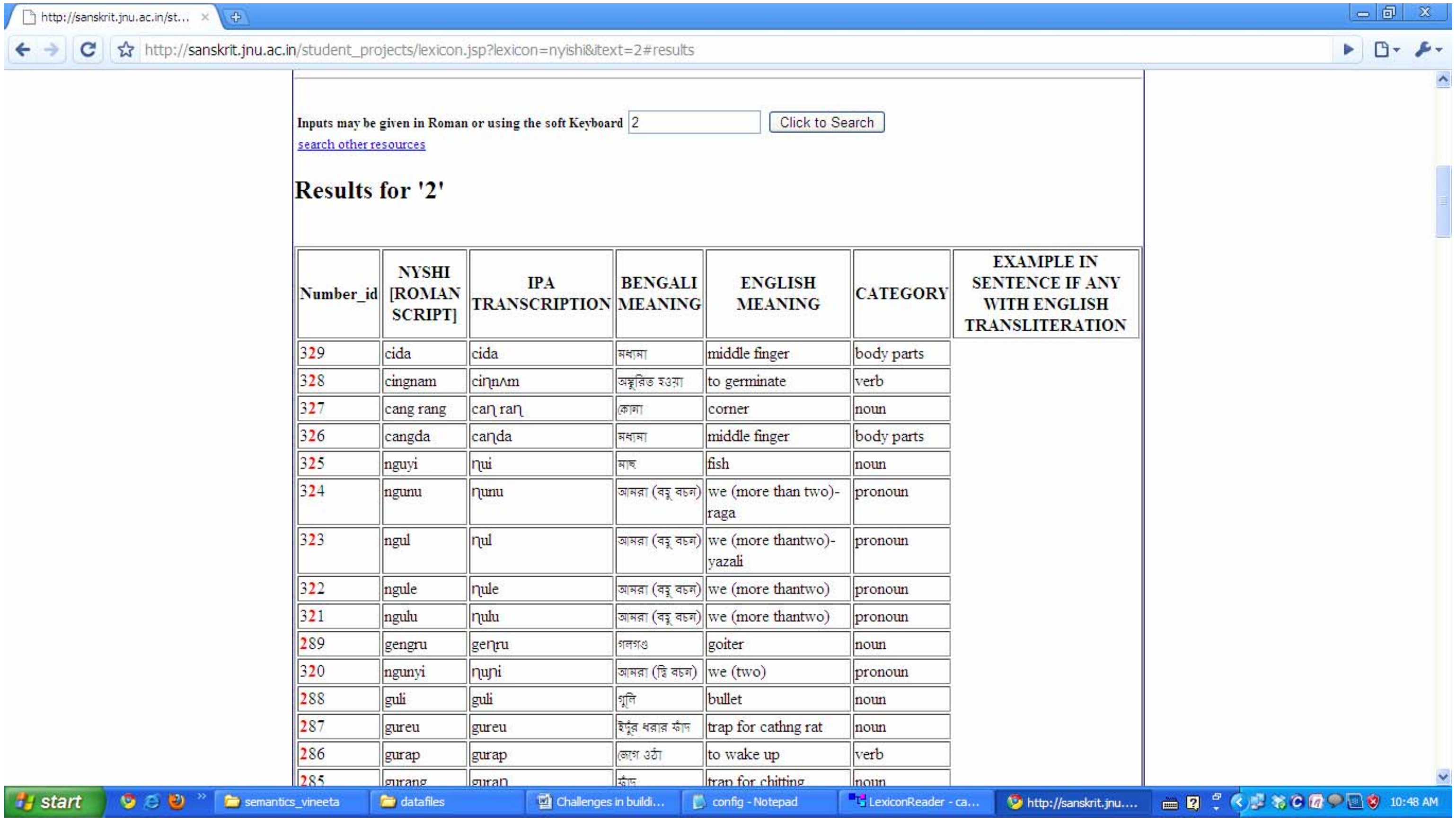
Home
Language Processing Tools
Lexical Resources
e-Learning
Corpora/e-Text
Dissertation
Feedback

***NYISHI-BENGALI-ENGLISH DICTIONARY*** **Project**

Author : *ATANU SAHA*
Center : *Center for Linguistics*
Course : *Multimedia storage and retrieval techniques for culture, Special Center for Sanskrit Studies,J.N.U.*
Semester : *Monsoon 2007*
Under the supervision of : *Dr. Girish Nath Jha*

*The World Of Nyishi......*

# Screen shot 2



Inputs may be given in Roman or using the soft Keyboard 2 Click to Search

search other resources

## Results for '2'

| Number_id | NYSHI [ROMAN SCRIPT] | IPA TRANSCRIPTION | BENGALI MEANING | ENGLISH MEANING | CATEGORY | EXAMPLE IN SENTENCE IF ANY WITH ENGLISH TRANSLITERATION |
|---|---|---|---|---|---|---|
| 329 | cida | cida | মধ্যমা | middle finger | body parts | |
| 328 | cingnam | ciŋnʌm | অঙ্কুরিত হওয়া | to germinate | verb | |
| 327 | cang rang | caŋ raŋ | কোনা | corner | noun | |
| 326 | cangda | caŋda | মধ্যমা | middle finger | body parts | |
| 325 | nguyi | ŋui | মাছ | fish | noun | |
| 324 | ngunu | ŋunu | আমরা (বহু বচন) | we (more than two)-raga | pronoun | |
| 323 | ngul | ŋul | আমরা (বহু বচন) | we (more thantwo)-yazali | pronoun | |
| 322 | ngule | ŋule | আমরা (বহু বচন) | we (more thantwo) | pronoun | |
| 321 | ngulu | ŋulu | আমরা (বহু বচন) | we (more thantwo) | pronoun | |
| 289 | gengru | geŋru | গলগণ্ড | goiter | noun | |
| 320 | ngunyi | ŋuɲi | আমরা (দ্বি বচন) | we (two) | pronoun | |
| 288 | guli | guli | গুলি | bullet | noun | |
| 287 | gureu | gureu | ইঁদুর ধরার ফাঁদ | trap for cathng rat | noun | |
| 286 | gurap | gurap | জেগে ওঠা | to wake up | verb | |
| 285 | gurang | guraŋ | ফাঁদ | trap for chitting | noun | |

# Challenges

- First of all it was a daunting task to collect the data because most of the people opine that the new generation is using this language less and they are switching to some other languages like Assamese, Hindi and English.

- They are more interested in learning Assamese which is the predominant language of that area or English which is always considered as a language of opportunity.

# Challenges …

- We had to keep mind that we are representing three languages at a time and therefore data should be organized and comprehensive to all the users.

- Though there are some short comings like the Unicode font problems etc.

# Challenges …

- We are not claiming that the font problem is fully resolved in our dictionary.

- Secondly to feed that data in a simple and representable way to the computer so that it understands it and produces the correct set of information.

# Challenges …

- We decided to save our jsp page in Unicode throughout the programming.

# Inference & Future prospects

- In our current lexicon there are more than 700 Nyishi words along with their transcription.
- More words and sentences need to collect.
- In this virtual dictionary data can be added anytime just by sending it to us.

# Future….

- As we all know India is a huge country with its rich heritage, culture and plenty of languages it's a novel way to archive and represent those prospects.
- Further with the given set of classes and the environment many more tribal culture and languages can be represented and will attract larger mass quite easily.

# Thank you ☺

atanu.jnu@gmail.com

girishjha@gmail.com

Questions !

Comments !!

Suggestions !!!