Odia resource grammar

By Bibek Behera
IIT Bombay
at
3rd GF summer school, Aug 18-31, 2013
Frauinsel, Bavaria, Germany

Content

- Introduction
- Motivation
- Current work at IIT Bombay
- Impact
  - Odia resource grammar
  - An Odia to 28 language MT system
- Things I have learnt
- Conclusion
- References

INTRODUCTION

Oriya language

- **Oriya** (ଓଡ଼ିଆ *oṛiā*), officially spelled **Odia** is an Indian language, belonging to the **Indo-Aryan branch** of the Indo-European language family.
- It is mainly spoken in the Indian states of Odisha and in parts of West Bengal, Jharkhand, Chhattisgarh and Andhra Pradesh.

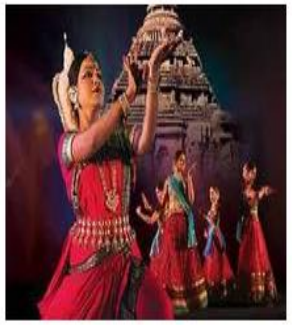
- 31 millions of people in India speak in this language in Orissa and also in some neighbour states.

Odia script

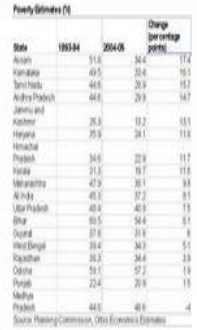
ଅ a
ଆ ā
ଇ i
ଈ ī
ଉ u
ଊ ū
ଋ ṛ
ଏ e
ଐ ai
ଓ o
ଔ au
କ ka
ଖ kha
ଗ ga
ଘ gha
ଙ ṅa
ଚ ca
ଛ cha
ଜ ja
ଝ jha
ଞ ña
ଟ ṭa
ଠ ṭha
ଡ ḍa
ଢ ḍha
ଣ ṇa
ତ ta
ଥ tha
ଦ da
ଧ dha
ନ na
ପ pa
ଫ pha
ବ ba
ଭ bha
ମ ma
ଯ ya
ର ra
ଲ la
ୱ va
ଶ śa
ଷ ṣa
ସ sa
ହ hā

କା kā
କି ki
କୀ kī
କୁ ku
କୂ kū
କୃ kṛ
କେ ke
କୈ kai
କୋ ko
କୌ kau

Pictures of Odisha



Poor Odisha



Poverty Estimates (%)

State	1993-94	2004-05	Change (percentage points)
Assam	51.8	34.4	17.4
Karnataka	49.5	33.4	16.1
Tamil Nadu	44.8	28.9	15.7
Andhra Pradesh	44.6	29.9	14.7
Jammu and Kashmir	25.3	13.2	12.1
Haryana	35.9	24.1	11.8
Himachal Pradesh	34.6	22.9	11.7
Kerala	31.3	19.7	11.6
Maharashtra	47.9	36.1	11.8
All India	45.3	37.2	8.1
Uttar Pradesh	40.4	40.9	7.5
Bihar	69.5	54.4	6.1
Gujarat	37.6	31.8	6
West Bengal	39.4	34.3	5.1
Rajasthan	38.2	34.4	3.8
Odisha	59.1	57.2	1.9
Punjab	22.4	20.9	1.5
Madhya Pradesh	44.6	48.9	-4

Source: Planning Commission, Orissa Economics Estimates

ORISSA POVERTY, CORPORATE PLUNDER AND RESISTANCE

Motivation

Social reason

- Orissa is one of the **poorest** states in India.
- But we have seen **development** with entry of technology.
- Odisha has **abundance** of natural beauty, but tourism has been **hardly promoted**.
- **Handicraft or handloom** like small scale industries exist but not well adveritised.
- GF applications in tourism and handicraft may improve the situation.

Motivation continued ...

Technical reason

- No translation system for Odia language
- Limited parallel corpus for Odia – English language.
- SMT system is not feasible right now.
- GF comes to rescue!

Current work at IIT Bombay

- **No work** in Odia to English translation.
- We work in statistical machine translation (SMT) such as phrase based machine translation (MT), factor based MT and hierarchical based MT for various **English to Indian** language translation.
- SMT systems **require huge parallel corpus** even for decently accurate systems.
- We are stuck because good quality parallel corpus for Odia English is not available.

IMPACT OF MY WORK

Programming assignments

- Some fun assignments:-
  - Reverse
    - Generated reverse of a string that has “a b c” in its lexicon.
    - Funny that no language is actually reverse of the other.
  - Catalan
    - generates catalan number of trees
      - Fish with fish with fish generates
        - » With Fish (With Fish Fish)
        - » With (With Fish Fish) Fish

Programming assignments ...

- Some serious work
  - Foods grammar in Odia (1-2 days)
    - *“ଏଇ ଛେନା ତାଜା ଅଟେ” - “this cheese is fresh”*
    - Restricted in the sense that it talks about food only
    - Also there are no verbs except the auxillary verb
    - Good for beginners
    - **Implemented on GF website**.
  - Mini resource grammar in Odia (5-6 days)
    - “ପ୍ରତ୍ୟେକ ସବୁଜ ଘରଟି ଛୋଟ ଅଟେ ଏବଂ ସିଏ ସେଇ ସବୁଜ ଘରଗୁଡିକ ପ୍ରେମ କରେ” – “every green house is small and she loves those green houses”.
    - **Implemented on GF website.**
    - Much more open than foods grammar.

Programming assignments...

- Mini resource grammar in Odia continued ...
  - Noun morphology is similar to that in english but even simpler because we have to **add a word** “ଗୁଡିକ” (gudika) like “ଘରଗୁଡିକ” (ghargudika) for plural.
  - Contains verbs, their morphologies as well as morphologies for the negation i.e. “ନୁହେ” (nuhe) (not) for singular and "ନୁହନ୍ତି“ (nuhanti) for plural.
  - In case of verbs, I have added about 10 forms although in theory there are thirteen forms but they overlap.
    - Various forms of the word walk -*chalibaa, chale, chalu, chala, chalanti, chalithili, chalithilu, chalithila, chalithilaa, chalithile.*

Programming assignments ...

- Mini RG continued ...
  - Some of the rules are listed below.
    - *VInf => chalibaa;*
    - *VPres Sg Per1 => chale;*
    - *VPres Sg Per2 => chalu;*
    - *VPres Sg Per3 => chale;*
    - *VPres Pl Per1 => chalu;*
  - Noun verb agreement has been handled.
  - Gender does not play a role in Odia morphology.
  - But still the grammar has limited lexicon.
    - Solution:- Move towards full resource grammar.

Things i have learnt from the camp

- Speech translation systems – Could be useful in many real life applications.
- Application of GF in controlled natural languages and semantic web.
- Applications like phrasebook, facebook were really interesting.
- Having food for 2 hours and digest properly.
- Drinking and walking around the Island.

Conclusions

- GF is a powerful tool to produce high precision translations.
- Social applications of GF as in tourism and small scale industries can make the technology more useful.
- In context of Indian languages, GF can play a crucial role to open those languages to Indians as well as foreigners.

References

- http://incredibleorissa.com/en/
- http://en.wikipedia.org/wiki/Oriya_language
- Link to GF webpage
- Link to IIT Bombay

Thanks = ଧନ୍ୟବାଦ (Dhanyabaad)