Language, technology, and identity

Developing linguistic resources in the framework of Cooperation for Development
1 INDIA
Jawaharlal Nehru University of New Delhi

2 BRASILE
Universidade Federal de Minas Gerais
4 PROJECTS ACTIVATED IN THE LAST 5 YEARS
4 DISSEMINATION EVENTS IN BRAZIL
3 UNIVERSITIES INVOLVED, 2 FROM DEVELOPING COUNTRIES
7 PHD STUDENTS INVOLVED, ALL FROM DEVELOPING COUNTRIES
14 RESEARCHERS INVOLVED, 3 FROM DILEF DEPARTMENT
8 LANGUAGES INCLUDED IN Multi-LINGUAL RESOURCES, EACH ONE SPOKEN IN DEVELOPING COUNTRIES
Languages of India in the IMAGACT Multilanguage and Multimodal Ontology of Actions

Jawaharlal Nehru University of New Delhi
IMAGACT Ontology of Action

- IMAGACT web application is an online dictionary delivered by DILEF (LABLITA research unit) that can be used for language teaching
- The core Ontology can be also exploited for the development of enabling technologies in the fields of artificial intelligence and human-machine interaction
- [http://imagact.lablita.it](http://imagact.lablita.it)
EN turn (10)

- John turns over
- Mary turns left
- John turns the paper over
- Mary turns the page
- Mary turns the handle
Languages of India

• With 29 states, 1,721 spoken languages, 22 of which are official languages of the country, India needs to face a hard work to manage its linguistic variety

• Although Hindi is the main Indian language (with over 400 millions native speaker), the majority of Indian people are mother-tongue speakers of other languages

• Innovative teaching programs, reciprocal understanding, and technology support must be guaranteed for each of these languages
IMAGACT4INDIA

• IMAGACT extension to Indian languages, developed in cooperation with the Jawaharlal Nehru University of New Delhi, aims to spread the usage of the IMAGACT web dictionary to a wide part of Indian people by adding the support for the major Indian languages.

• Currently Hindi and Urdu are fully implemented in the resource and other 7 languages are under development, covering about 700 millions of people.
Mary turns left

John turns the paper over
Other Indian languages in development: Assamese, Bengali, Magahi, Manipuri, Odia, Tamil, Sanskrit

The expected impact of this Project is high, given the huge number of potential users belonging to linguistic communities for which multilingual web resources are not easily available
<table>
<thead>
<tr>
<th>Corpus verbs</th>
<th>Type</th>
<th>Lang.</th>
<th>Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>allontanare</td>
<td>PRO</td>
<td>IT</td>
<td>Marta si allontana dal tavolo</td>
</tr>
<tr>
<td>move</td>
<td>INST</td>
<td>EN</td>
<td>Mary moves away from the table</td>
</tr>
<tr>
<td>spostare</td>
<td>INST</td>
<td>IT</td>
<td>Marta si sposta</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verb</th>
<th>Transliteration</th>
<th>Rejected</th>
<th>Lang.</th>
<th>Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>to distance</td>
<td></td>
<td></td>
<td>EN</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verb</th>
<th>Transliteration</th>
<th>Lang.</th>
<th>Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>দূরে যেবার</td>
<td>dUre sarA</td>
<td>Ba</td>
<td>গীতা তেকিলটি থেকে দূরে যাচ্ছে</td>
</tr>
<tr>
<td>দূরে যেবার যাচ্ছে</td>
<td>dUre sare yAoYA</td>
<td>Ba</td>
<td>গীতা তেকিলটি থেকে দূরে যাচ্ছে</td>
</tr>
<tr>
<td>দূরে যাচ্ছে</td>
<td>dUre yAoYA</td>
<td>Ba</td>
<td>গীতা তেকিলটি থেকে দূরে যাচ্ছে</td>
</tr>
<tr>
<td>দূরে চলে যাচ্ছে</td>
<td>dUre chale yAoYA</td>
<td>Ba</td>
<td>গীতা তেকিলটি থেকে দূরে চলে যাচ্ছে</td>
</tr>
</tbody>
</table>
Language and identity: spontaneous spoken Brazilian Portuguese resources in the Romance context

Universidade Federal de Minas Gerais
Spoken language resources

• Within the agreement between UNIFI and UFMG (Belo Horizonte, Brazil), a large corpus of spoken Brazilian Portuguese has been released

• C-ORAL-BRASIL, built on the model of LABELITA resources
  • C-ORAL-ROM corpora of Romance languages: French, Italian, Portuguese, and Spanish
  • Standard framework for speech representation
A part of the C-ORAL-BRASIL corpus has been also extensively annotated and is a part of the DB-IPIC database.

- DataBase for Information Patterning Interlinguistic Comparison, maintained by LABLITA
- [http://www.lablita.it/app/dbipic/](http://www.lablita.it/app/dbipic/)
Spoken Language & Identity

• C-ORAL-BRASIL is fully comparable with the Portuguese C-ORAL-ROM resource, and represents the specificities of the Brazilian variety of spoken Portuguese
  • Vs. a substantial homogeneity between the Brazilian and the European written varieties

• This resource will give an important contribution to the affirmation of the Brazilian autonomous linguistic and cultural identity