POS Tagging ILCI-Telugu Corpus

S.Arulmozi
Dravidian University

Overview

- Indian Languages Corpora Initiative
- Telugu Corpus
- POS Annotation
- Issues

ILCI

- The Indian Languages Corpora Initiative (ILCI) is a **research project for technology development for Indian languages.**
- Special Centre for Sanskrit Studies of **Jawaharlal Nehru University** is coordinating this national project and **is the consortium leader of the ILCI project**.

Consortium Members

- Punjabi University for Punjabi
- JNU (Center for Indian languages) for Urdu
- ISI Kolkata for Bangla
- Utkal University for Oriya
- IIT Mumbai for Marathi
- Gujarat University for Gujarati
- **Dravidian University for Telugu**
- Tamil University for Tamil
- IITM-K Trivandrum for Malayalam
- Goa University for Konkani
- **Each consortium member will develop corpora and standards in their respective languages.**

- The **main objective**
  - **to build an annotated parallel corpora** (Hindi to 11 Indian languages along with English) **with standards** for 12 major Indian languages including English in the **domain of tourism and health**.
- Major aims of the project are
  - to evolve draft standards
  - build parallel corpora in the domain of tourism and health (Hindi-English and Hindi-Indian languages) and
  - annotate (label) the parallel corpora.

- Evolving Draft Standards includes evaluation of existing corpora and tools that have been developed as part of various projects under Technology Development in Indian Languages (TDIL), and evaluating existing standards for their usability.
- Standards for corpora collection, for corpora encoding and for corpora validation
- The task of Corpora development includes corpora collection in Hindi, parallel corpora in 11 Indian languages and parallel corpora in English.

- The basic starting point for this project is a list of 50,000 Hindi sentences used in the tourism and health domain.
- A list of data source institutions including Tourism and Health departments was made to collect data for Hindi.
- Parallel aligned corpora with Hindi as source language in the given 11 Indian languages and English has been created as per the standards evolved.
- Annotated corpora in these 11 Indian languages and English are in process as per the BIS standards (Feb 2012 deadline)

Telugu ILCI Corpus

- 50 K sentences from Hindi into Telugu were translated
- 25 k each in tourism and health domain
- Annotation work is in progress (Feb 2012 deadline) based on BIS-POS Tagset (for Telugu)

Telugu Tagset - Noun

Common – N_NN	abbAyi, puswakaM
Proper – N_NNP	kqRNudu, rAvu, reVddi
Nloc – N_NST	munupu, kiMxata, moVxata, akkada

Pronoun

• Personal - PR__PRP	wanu, nenu, nuvvu
• Reflexive - PR__PRF	svayaM
• Relative - PR__PRL	eVvadu, exi eVppudu, eVkkada
• Reciprocal - PR__PRC	parasparaM
• Wh-word - PR__PRQ	eVvadu, exi

Demonstratives

- Deictic - DM__DMD A, I
- Relative - DM__DMR e
- Wh-word - DM__DMQ e

Verb

- Main - V__VM ammAnu, koVttanu, le,
- Finite - V__VM__VF ammAnu, koVttanu, le,
- Non-finite - V__VM__VNF wini, wine winaka
- Infinitive - V__VM__VINF wina, vina, ceVppa
- Gerund - V__VM__VNG winadaM, vinadaM
- Auxiliary - V__VAUX uMxi, kAxu, lexu, padu

- Adjective – JJ cakkani, maMci, peVxxa
- Adverb – RB veVMt ane, wvaragA
- Postposition - PSP ku, cewa, lona,bayata

Conjunction

- Coordinator - CC__CCD mariyu, iMkA , kAnI
- Subordinator - CC__CCS ani, aMte, aMtU
  - – Quotative - CC__CCS__UT ani
  - – Identifier – CC_CCS_ID aMte, aMtU

Particles

Default	RP__RPD	ayiwe,kUdA, mAwraM
Classifier	RP__CL	maMxi, guru
Interjection	RP__INJ	are, ayyo, oy,ammo
Intensifier	RP__INTF	cAla, eVnni, eVMwo
Negation	RP__NEG	wappa,vinA

Quantifiers

- General QT__QTF koVxxi, cAlA, koVMceVM
- Cardinals QT__QTC oVkati, reVMdu, mUdu
- Ordinals QT__QTO oVkato, reVMdava, mUdava

Residuals

- Foreign word RD__RDF A word written in a script other than the script of the current text
- Symbol RD__SYM
- Punctuation RD__PUNC
- Unknown RD__UNK
- Echowords RD__ECH

Issues

- 14653 prajalu ceppedEmanagA
- 23016 nammabaDinaDEmiTaMTE
- 14568 vATannIMTIlO
- 47 74 oka rakamaina vairas
- 23103 iMtEgAka
- 14699 binna-binna

Some Examples

- htd10582
- दिल्ली में इस अनुमति-पत्र को गवर्नमेंट ऑफ अरुणाचल , कौटिल्य मार्ग , चाणक्यपुरी से प्राप्त किया जा सकता है ।
- ఢిల్లీలో ఈ అనుమతి-పత్రాన్ని గవర్నమెంట్ ఆఫ్ అరుణాచల్, కౌటిల్యమార్గ్, చాణక్యపురి నుండి లభించడం జరుగుతుంది.
- DhillIlO I anumati-patrAnni gavarnameMT AP aruNAcal, kauTilyamArg, cANakyapuri nuMDi laBiMcaDaM jarugutuMdi.

- htd10626
- वैसे न्यू ’बोगाईगाँव’ से ’तेजू’ फिर जीप से व पैदल यात्रा कर परशुराम कुंड पहुँचा जा सकता है ।
- అలాగే న్యూ ’బోగైగ్రామ’ నుండి ’తెజూ’ మరళ జీఫ్తో మరియు కాళినడక ప్రయాణంచేసి పరుశురామ కుంఢ చేరుకోవచ్చు.
- alAgE nyU 'bOgaigrAma' nuMDi 'tejU' maraLa jIPtO mariyu kALi naDaka prayANaMcEsi paruSurAma kuMDha cErukOvaccu.

Acknowledgements

- BIS Tagset – Prof. G.Uma Maheswar Rao
- ILCI Consortia, DIT

Thank you!