

---

---

**IJCNLP 2008**

---

---

**The 6th Workshop on  
Asian Language Resources  
(ALR 6)**

Proceedings of the Workshop

11-12 January 2008  
Indian School of Business, Hyderabad, India

©2008 Asian Federation of Natural Language Processing

## **Sponsor**

Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sport, Science and Technology, MEXT Japan.

## Preface

This volume contains the papers presented at the sixth workshop on Asian Language Resources, held on 11–12 January 2008 in conjunction with the third International Joint Conference on Natural Language Processing (IJCNLP 2008).

Language resources have played an essential role in empirical approaches to natural language processing (NLP) for the last two decades. Previous concerted efforts on construction of language resources, particularly in the US and European countries, have laid a solid foundation for the pioneering NLP researches in these two communities. In comparison, the availability and accessibility of many Asian language resources are still very limited except for a few languages. Moreover, there is a greater diversity in Asian languages with respect to character sets, grammatical properties and the cultural background.

Motivated by such a context, we have organised a series of workshops on Asian language resources since 2001. This workshop series has contributed to the activation of the NLP research in Asia particularly of building and utilising corpora of various types and languages. In this sixth workshop, we had 31 submissions encompassing 13 languages. The paper selection was highly competitive compared with the last five workshops. The program committee selected 10 regular papers, 3 short papers and 8 resource reports for presentation at the workshop.

The workshop is comprised of two parts, technical sessions and a session devoted to reporting activities related to language resources in several languages. Following the resource report session, we have an open discussion on the collaboration in building, standardising and exchanging language resources in Asia. We hope this workshop further accelerates the already thriving NLP research in Asia.

Chu-Ren Huang  
Mikami Yoshiki  
*Workshop Co-chairs*

Hasida Kôiti  
Tokunaga Takenobu  
*Program Co-chairs*

## Organiser

### Workshop chairs

Huang, Chu-Ren *Academia Sinica*  
Mikami, Yoshiki *Nagaoka University of Technology*

### Program chairs

Hasida, Kôiti *National Institute of Advanced Industrial Science and Technology*  
Tokunaga, Takenobu *Tokyo Institute of Technology*

### Program Committee

Bhattacharyya, Pushpak *IIT, Bombay*  
Fang, Alex Chengyu *City University of Hong Kong*  
Riza, Hammam *IPTEKnet–BPPT*  
Hasida, Kôiti *National Institute of Advanced Industrial Science and Technology*  
He, Tingting *Huazhong Normal University*  
Huang, Chu-Ren *Academia Sinica*  
Hussain, Sarmad *National University of Computer & Emerging Sciences*  
Itahashi, Shuichi *National Institute of Informatics*  
Lu, Qin *Hong Kong Polytechnic University*  
Luong, Chi Mai *National Center for Sciences and Technologies of Vietnam*  
Mikami, Yoshiki *Nagaoka University of Technology*  
Nandasara, Shakrange Turrance *University of Colombo, School of Computing*  
Nguyen, Thi Minh Huyen *Hanoi University of Sciences*  
Oo, Thein *Myanmar Computer Federation*  
Rau, Victoria *Providence University*  
Rim, Hae-Chang *Korea University*  
Roxas, Rachel Edita O *De La Salle University, Manila*  
Shirai, Kiyooki *Japan Advanced Institute of Science and Technology*  
Somnertlamvanich, Virach *Thai Computational Linguistics Laboratory, NICT*  
Sui, Zhifang *Peking University*  
Tokunaga, Takenobu *Tokyo Institute of Technology*  
Vikas, Om *Indian Institute of Information Technology and Management*  
Zhao, Jun *Chinese Academy of Sciences*

This workshop is supported by Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sport, Science and Technology, MEXT Japan.

# Workshop Program

11-12 January 2008

Indian School of Business, Hyderabad, India

## Day 1 (11 January)

- 9:00 *Registration*
- 9:20 *Opening*
- 9:30 *Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems*  
Asif Ekbal and Sivaji Bandyopadhyay
- 9:55 *Gazetteer Preparation for Named Entity Recognition in Indian Languages*  
Sujan Kumar Saha, Sudeshna Sarkar and Pabitra Mitra
- 10:20 *Preliminary Chinese Term Classification for Ontology Construction*  
Gaoying Cui, Qin Lu and Wenjie Li
- 10:45 *Break*
- 11:05 *Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms*  
Makiko Matsuda, Tomoe Takahashi, Hiroki Goto, Yoshikazu Hayase, Robin Lee Nagano and Yoshiki Mikami
- 11:30 *Selection of XML tag set for Myanmar National Corpus*  
Wunna Ko Ko and Thin Zar Phyo
- 11:55 *Myanmar Word Segmentation using Syllable level Longest Matching*  
Hla Hla Htay and Kavi Narayana Murthy
- 12:20 *Lunch*
- 13:50 *The Link Structure of Language Communities and its Implication for Language-specific Crawling*  
Rizza Caminero and Yoshiki Mikami
- 14:15 *A Multilingual Multimedia Indian Sign Language Dictionary Tool*  
Tirthankar Dasgupta, Sambit Shukla, Sandeep Kumar, Synny Diwakar and Anupam Basu
- 14:40 *A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus*  
Deniz Zeyrek and Bonnie Webber
- 15:05 *Towards an Annotated Corpus of Discourse Relations in Hindi*  
Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi
- 15:30 *Break*
- 15:50 *A Semantic Study on Yami Ontology in Traditional Songs*  
Yin-Sheng Tai, D. Victoria Rau and Meng-Chien Yang
- 16:05 *Assessment and Development of POS Tag Set for Telugu*  
Rama Sree R.J., Uma Maheswara Rao G and Madhu Murthy K.V.
- 16:20 *Designing a Common POS-Tagset Framework for Indian Languages*  
Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L and Subbarao K V

## **Day 2 (12 January)**

- 9:00 *Resources Report on Languages of Indonesia*  
Hammam Riza
- 9:15 *Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List*  
Ryo Nishimura, Yasuhiko Watanabe and Yoshihiro Okada
- 9:30 *Corpus building for Mongolian language*  
Purev Jaimai and Odbayar Chimeddorj
- 9:45 *Resources for Urdu Language Processing*  
Sarmad Hussain
- 10:00 *Balanced Corpus of Contemporary Written Japanese*  
Kikuo Maekawa
- 10:15 *Break*
- 10:35 *A Basic Framework to Build a Test Collection for the Vietnamese Text Categorization*  
Viet Hoang-Anh, Thu Dinh-Thi-Phuong and Thang Huynh-Quyet
- 10:50 *Enhanced Tools for Online Collaborative Language Resource Development*  
Virach Sornlertlamvanich, Thatsanee Charoenporn, Suphanut Thayaboon, Chumpol Mokarat and Hitoshi Isahara
- 11:05 *Japanese Effort Toward Sharing Text and Speech Corpora*  
Shuichi Itahashi and Kôiti Hasida
- 11:20 *Open Discussion*
- 12:20 *Closing*

## Table of Contents

### ⟨Regular papers⟩

<i>Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems</i> Asif Ekbal and Sivaji Bandyopadhyay .....	1
<i>Gazetteer Preparation for Named Entity Recognition in Indian Languages</i> Sujan Kumar Saha, Sudeshna Sarkar and Pabitra Mitra .....	9
<i>Preliminary Chinese Term Classification for Ontology Construction</i> Gaoying Cui, Qin Lu and Wenjie Li .....	17
<i>Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms</i> Makiko Matsuda, Tomoe Takahashi, Hiroki Goto, Yoshikazu Hayase, Robin Lee Nagano and Yoshiki Mikami .....	25
<i>Selection of XML tag set for Myanmar National Corpus</i> Wunna Ko Ko and Thin Zar Phyo .....	33
<i>Myanmar Word Segmentation using Syllable level Longest Matching</i> Hla Hla Htay and Kavi Narayana Murthy .....	41
<i>The Link Structure of Language Communities and its Implication for Language-specific Crawling</i> Rizza Caminero and Yoshiki Mikami .....	49
<i>A Multilingual Multimedia Indian Sign Language Dictionary Tool</i> Tirthankar Dasgupta, Sambit Shukla, Sandeep Kumar, Synny Diwakar and Anupam Basu .....	57
<i>A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus</i> Deniz Zeyrek and Bonnie Webber .....	65
<i>Towards an Annotated Corpus of Discourse Relations in Hindi</i> Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi .....	73

### ⟨Short papers⟩

<i>A Semantic Study on Yami Ontology in Traditional Songs</i> Yin-Sheng Tai, D. Victoria Rau and Meng-Chien Yang .....	81
<i>Assessment and Development of POS Tag Set for Telugu</i> Rama Sree R.J., Uma Maheswara Rao G and Madhu Murthy K.V. ....	85
<i>Designing a Common POS-Tagset Framework for Indian Languages</i> Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L and Subbarao K V .....	89

### ⟨Resource reports⟩

<i>Resources Report on Languages of Indonesia</i> Hammam Riza .....	93
<i>Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List</i> Ryo Nishimura, Yasuhiko Watanabe and Yoshihiro Okada .....	95

<i>Corpus building for Mongolian language</i> Purev Jaimai and Odbayar Chimeddorj .....	97
<i>Resources for Urdu Language Processing</i> Sarmad Hussain .....	99
<i>Balanced Corpus of Contemporary Written Japanese</i> Kikuo Maekawa .....	101
<i>A Basic Framework to Build a Test Collection for the Vietnamese Text Categorization</i> Viet Hoang-Anh, Thu Dinh-Thi-Phuong and Thang Huynh-Quyet .....	103
<i>Enhanced Tools for Online Collaborative Language Resource Development</i> Virach Sornlertlamvanich, Thatsanee Charoenporn, Suphanut Thayaboon, Chumpol Mokarat and Hitoshi Isahara .....	105
<i>Japanese Effort Toward Sharing Text and Speech Corpora</i> Shuichi Itahashi and Kôiti Hasida .....	107



# Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems

**Asif Ekbal**

Department of Computer Science and  
Engineering, Jadavpur University  
Kolkata-700032, India  
asif.ekbal@gmail.com

**Sivaji Bandyopadhyay**

Department of Computer Science and  
Engineering, Jadavpur University  
Kolkata-700032, India  
sivaji\_cse\_ju@yahoo.com

## Abstract

The rapid development of language tools using machine learning techniques for less computerized languages requires appropriately tagged corpus. A Bengali news corpus has been developed from the web archive of a widely read Bengali newspaper. A web crawler retrieves the web pages in Hyper Text Markup Language (HTML) format from the news archive. At present, the corpus contains approximately 34 million wordforms. The *date*, *location*, *reporter* and *agency tags* present in the web pages have been automatically named entity (NE) tagged. A portion of this partially NE tagged corpus has been manually annotated with the sixteen NE tags with the help of *Sanchay Editor*<sup>1</sup>, a text editor for Indian languages. This NE tagged corpus contains 150K wordforms. Additionally, 30K wordforms have been manually annotated with the twelve NE tags as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages<sup>2</sup>. A table driven semi-automatic NE tag conversion routine has been developed in order to convert the sixteen-NE tagged corpus to the twelve-NE tagged corpus. The 150K NE tagged corpus has been used to develop Named Entity Recognition (NER) system in Bengali using pattern directed shallow parsing approach, Hidden Markov Model (HMM), Maximum Entropy (ME) Model, Condi-

tional Random Field (CRF) and Support Vector Machine (SVM). Experimental results of the 10-fold cross validation test have demonstrated that the SVM based NER system performs the best with an overall F-Score of 91.8%.

## 1 Introduction

The mode of language technology work has been changed dramatically since the last few years with the web being used as a data source in a wide range of research activities. The web is anarchic, and its use is not in the familiar territory of computational linguistics. The web walked into the ACL meetings started in 1999. The use of the web as a corpus for teaching and research on language technology has been proposed a number of times (Rundel, 2000; Fletcher, 2001; Robb, 2003; Fletcher, 2003). There is a long history of creating a standard for western language resources. The human language technology (HLT) society in Europe has been particularly zealous for the standardization, making a series of attempts such as EAGLES<sup>3</sup>, PROLE/SIMPLE (Lenci et al., 2000), ISLE/MILE (Calzolari et al., 2003; Bertagna et al., 2004) and more recently multilingual lexical database generation from parallel texts in 20 European languages (Giguet and Luquet, 2006). On the other hand, in spite of having great linguistic and cultural diversities, Asian language resources have received much less attention than their western counterparts. A new project (Takenobou et al., 2006) has been started to create a common standard for Asian language resources. They have extended an existing description framework, the

---

<sup>1</sup> [sourceforge.net/project/nlp-sanchay](http://sourceforge.net/project/nlp-sanchay)

<sup>2</sup> <http://ltrc.iit.ac.in/ner-ssea-08>

---

<sup>3</sup> <http://www.ilc.cnr.it/Eagles96/home.html>

MILE (Bertagna et al., 2004), to describe several lexical entries of Japanese, Chinese and Thai. India is a multilingual country with the enormous cultural diversities. (Bharati et al., 2001) reports on efforts to create lexical resources such as transfer lexicon and grammar from English to several Indian languages and dependency tree bank of annotated corpora for several Indian languages. Corpus development work from web can be found in (Ekbal and Bandyopadhyay, 2007d) for Bengali.

Named Entity Recognition (NER) is one of the core components in most Information Extraction (IE) and Text Mining systems. During the last few years, the probabilistic machine learning methods have become state of the art for NER (Bikel et al., 1999; Chieu and Ng, 2002) and for field extraction (McCallum et al., 2000). Most prominently, Hidden Markov Models (HMMs) have been used for the information extraction task (Bikel et al., 1999). Beside HMM, there are other systems based on Support Vector Machine (Sun et al., 2003) and Naïve Bayes (De Sitter and Daelemans, 2003). Maximum Entropy (ME) conditional models like ME Markov models (McCallum et al., 2000) and Conditional Random Fields (CRFs) (Lafferty et al., 2001) were reported to outperform the generative HMM models on several IE tasks.

The existing works in the area of NER are mostly in non-Indian languages. There has been a very little work in the area of NER in Indian languages (ILs). In ILs, particularly in Bengali, the work in NER can be found in (Ekbal and Bandyopadhyay, 2007a; Ekbal and Bandyopadhyay, 2007b; Ekbal et al., 2007c). Other than Bengali, the work on NER can be found in (Li and McCallum, 2003) for Hindi.

Newspaper is a huge source of readily available documents. In the present work, the corpus has been developed from the web archive of a very well known and widely read Bengali newspaper. Bengali is the seventh popular language in the world, second in India and the national language of Bangladesh. A code conversion routine has been written that converts the proprietary codes used in the newspaper into the standard Indian Script Code for Information Interchange (ISCII) form, which can be processed for various tasks. A separate code conversion routine has been developed for converting ISCII codes to UTF-8 codes. A portion of this corpus has been manually annotated with the sixteen NE tags as described in Table 3. Another por-

tion of the corpus has been manually annotated with the twelve NE tags as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages. A table driven semi-automatic NE tag conversion routine has been developed in order to convert this corpus to a form tagged with the twelve NE tags. The NE tagged corpus has been used to develop Named Entity Recognition (NER) system in Bengali using pattern directed shallow parsing approach, HMM, ME, CRF and SVM frameworks.

A number of detailed experiments have been conducted to identify the best set of features for NER in Bengali. The ME, CRF and SVM based NER models make use of the language independent as well as language dependent features. The language independent features could be applicable for NER in other Indian languages also. The system has demonstrated the highest F-Score value of 91.8% with the SVM framework. One possible reason behind its best performance may be the flexibility of the SVM framework to handle the morphologically rich Indian languages.

## 2 Development of the Named Entity Tagged Bengali News Corpus

### 2.1 Language Resource Acquisition

A web crawler has been developed that retrieves the web pages in Hyper Text Markup Language (HTML) format from the news archive of a leading Bengali newspaper within a range of dates provided as input. The crawler generates the Universal Resource Locator (URL) address for the index (first) page of any particular date. The index page contains actual news page links and links to some other pages (e.g., Advertisement, TV schedule, Tender, Comics and Weather etc.) that do not contribute to the corpus generation. The HTML files that contain news documents are identified and the rest of the HTML files are not considered further.

### 2.2 Language Resource Creation

The HTML files that contain news documents are identified by the web crawler and require cleaning to extract the Bengali text to be stored in the corpus along with relevant details. The HTML file is scanned from the beginning to look for tags like `<fontFACE=BENGALI_FONT_NAME>...<font>`, where the BENGALI\_FONT\_NAME is the name

of one of the Bengali font faces as defined in the news archive. The Bengali text enclosed within font tags are retrieved and stored in the database after appropriate tagging. Pictures, captions and tables may exist anywhere within the actual news. Tables are integral part of the news item. The pictures, its captions and other HTML tags that are not relevant to our text processing tasks are discarded during the file cleaning. The Bengali news corpus has been developed in both ISCII and UTF-8 codes. The tagged news corpus contains 108,305 number of news documents with about five (5) years (2001-2005) of news data collection. Some statistics about the tagged news corpus are presented in Table 1.

Total number of news documents in the corpus	108, 305
Total number of sentences in the corpus	2, 822, 737
Average number of sentences in a document	27
Total number of wordforms in the corpus	33, 836, 736
Average number of wordforms in a document	313
Total number of distinct wordforms in the corpus	467, 858

Table 1. Corpus Statistics

### 2.3 Language Resource Annotation

The Bengali news corpus collected from the web is annotated using a tagset that includes the type and subtype of the news, title, date, reporter or agency name, news location and the body of the news. A part of this corpus is then tagged with a tagset, consisting of sixteen NE tags and one non-NE tag. Also, a part of the corpus has been tagged with a tagset of twelve NE tags<sup>4</sup>, defined for the IJCNLP-08 NER Shared Task for South and South East Asian Languages.

A news corpus, whether in Bengali or in any other language, has different parts like title, date, reporter, location, body etc. To identify these parts in a news corpus the tagset, described in Table 2, has been defined. The reporter, agency, location, date, bd, day and ed tags help to recognize the person name, organization name, location name

and the various date expressions that appear in the fixed places of the newspaper. These tags are not able to recognize the various NEs that appear within the actual news body.

In order to identify NEs within the actual news body, we have defined a tagset consisting of seventeen tags. We have considered the major four NE classes, namely ‘Person name’, ‘Location name’, ‘Organization name’ and ‘Miscellaneous name’. Miscellaneous names include the date, time, number, percentage and monetary expressions. The four major NE classes are further divided in order to properly denote each component of the multiword NEs. The NE tagset is shown in Table 3 with the appropriate examples.

We have also tagged a portion of the corpus as part of the IJCNLP-08 NER Shared Task for South and South East Asian Languages. This tagset has twelve different tags. The underlying reason for adopting these tags was the necessity of a slightly finer tagset for various natural language processing (NLP) applications and particularly for machine translation. The IJCNLP-08 NER shared task tagset is shown in Table 4.

One important aspect of IJCNLP-08 NER shared task was to annotate only the maximal NEs and not the structures of the entities. For example, *mahatma gandhi road* is annotated as location and assigned the tag ‘NEL’ even if *mahatma* and *gandhi* are NE title person and person name, respectively, according to the IJCNLP-08 shared task tagset. These internal structures of the entities need to be identified during testing. So, *mahatma gandhi road* will be tagged as *mahatma*/NETP *gandhi*/NEP *road*/NEL. The structure of the tagged element using the *Shakti Standard Format* (SSF)<sup>5</sup> will be as follows:

```

1      ((      NP      <ne=NEL>
1.1    ((      NP      <ne=NEP>
1.1.1  ((      NP      <ne=NETP>
1.1.1.1 mahatma
          ))
1.1.2  gandhi
          ))
1.2    road
          ))

```

<sup>4</sup><http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=3>

<sup>5</sup><http://shiva.iiit.ac.in/SPSAL 2007/ssf.html>

## 2.4 Partially Tagged News Corpus Development

A news document is stored in the corpus in XML format using the tagset, mentioned in Table 2. In the HTML news file, the date is stored at first and is divided into three parts. The first one is the date according to Bengali calendar, second one is the day in Bengali and the last one is the date according to English calendar. Both Bengali and English

dates are stored in the form ‘day month year’.

A sequence of four Bengali digits separates the Bengali date from the Bengali day. The English date starts with one/two digits in Bengali font. Bengali date, day and English date can be distinguished by checking the appearance of the numerals and these are tagged as <bd>, <day> and <ed>, respectively. For e.g., *25 sraban 1412 budhbar 10 august 2005* is tagged as shown in Table 5.

Tag	Definition	Tag	Definition	Tag	Definition
header	Header of the news documents	day	Day	body	Body of the news document
title	Headline of the news document	ed	English date	p	Paragraph
t1	1st headline of the title	reporter	Reporter name	table	Information in tabular form
t2	2nd headline of the title	agency	Agency providing news	tc	Table column
date	Date of the news document	location	News location	tr	Table row
bd	Bengali date				

Table 2. News Corpus Tagset

Tag	Meaning	Example
PER	Single word person name	<i>sachin</i> / PER, <i>manmohan</i> /PER
LOC	Single word location name	<i>jadavpur</i> / LOC, <i>delhi</i> /LOC
ORG	Single word organization name	<i>infosys</i> / ORG, <i>tifr</i> /ORG
MISC	Single word miscellaneous name	<i>100%</i> / MISC, <i>100</i> /MISC
B-PER I-PER E-PER	Beginning, Internal or the end of a multiword person name	<i>sachin</i> / B-PER <i>ramesh</i> / I-PER <i>tendulkar</i> / E- PER
B-LOC I-LOC E-LOC	Beginning, Internal or the end of a multiword location name	<i>mahatma</i> / B-LOC <i>gandhi</i> / I-LOC <i>road</i> / E-LOC
B-ORG I-ORG E-ORG	Beginning, Internal or the end of a multiword organization name	<i>bhaba</i> / B-ORG <i>atomic</i> / I-ORG <i>research</i> / I-ORG <i>centre</i> / E-ORG
B-MISC I-MISC E-MISC	Beginning, Internal or the end of a multiword miscellaneous name	<i>10 e</i> / B-MISC <i>magh</i> / I-MISC <i>1402</i> / E-MISC
NNE	Words that are not named entities	<i>neta</i> /NNE, <i>bidhansabha</i> /NNE

Table 3. Named Entity Tagset

NE tag	Meaning	Example
NEP	Person name	<i>sachin ramesh tendulkar</i> / NEP
NEL	Location name	<i>mahatma gandhi road</i> / NEL
NEO	Organization name	<i>bhaba atomic research centre</i> / NEO
NED	Designation	<i>chairman</i> /NED, <i>sangsad</i> /NED
NEA	Abbreviation	<i>b a</i> /NEA, <i>c m d a</i> /NEA, <i>b j p</i> /NEA
NEB	Brand	<i>fanta</i> /NEB, <i>windows</i> /NEB
NETP	Title-person	<i>sriman</i> /NED, <i>sree</i> /NED
NETO	Title-object	<i>american beauty</i> /NETO
NEN	Number	<i>10</i> /NEN, <i>dash</i> /NEN
NEM	Measure	<i>tin din</i> /NEM, <i>panch keji</i> /NEM
NETE	Terms	<i>hidden markov model</i> /NETE
NETI	Time	<i>10 e magh 1402</i> /NETI

Table 4. IJCNLP-08 NER Shared Task Tagset

Original date pattern	Tagged date pattern
	<date>
<i>25 sraban 1412</i>	<bd> <i>25 sraban 1412</i> </bd>
<i>budhbar</i>	<day> <i>budhbar</i> </day>
<i>10 august 2005</i>	<ed> <i>10 august 2005</i> </ed>
	</date>

Table 5. Example of a Tagged Date Pattern

## 2.5 Named Entity Tagged Corpus Development

The partially NE tagged corpus contains 34 million wordforms and are in both ISCII and UTF-8 forms. A portion of this corpus, containing 150K wordforms, has been manually annotated with the sixteen NE tags that are listed in Table 3. The corpus has been annotated with the help of *Sanchay Editor*, a text editor for Indian languages. The detailed statistics of this NE-tagged corpus are given in Table 6. The corpus is in SSF form, which has the following structure:

```
<Story id="">
<Sentence id="">
1      biganni NNE
2      newton PER
3      .
</Sentence id="">
</Story id="">
```

Another portion of the partially NE tagged Bengali news corpus has been manually annotated as part of the IJCNLP-08 NER shared task with the twelve NE tags, as listed in Table 4. The annotation process has been very difficult due to the presence of a number of ambiguous NE tags. The potential ambiguous NE tags are: NED vs NETP, NEO vs NEB, NETE vs NETO, NETO vs NETP and NEN vs NEM. For example, it is difficult to decide whether ‘Agriculture’ is ‘NETE’, and if no then whether ‘Horticulture’ is ‘NETE’ or not. In fact, this the most difficult class to identify. This NE tagged corpus contains approximately 30K wordforms. Details statistics of this tagged corpus are shown in Table 7. This NE tagged corpus is in the following SSF form.

```
<Story id="">
<Sentence id="">
1  ((      NP      <ne=NEP>
1.1 ((      NP      <ne=NED>
1.1.1 biganni
      ))
1.1.2 newton NEP
      ))
2  .
</Sentence id="">
</Story id="">
```

NE Class	Number of wordforms	Number of distinct wordforms
Person name	20,455	15,663
Location name	11,668	5,579
Organization name	963	867
Miscellaneous name	11,554	3,227

Table 6. Statistics of the 150K-tagged Corpus

## 2.6 Tag Conversion

A tag conversion routine has been developed in order to convert the sixteen-NE tagged corpus of 150K wordforms to the corpus, tagged with the IJCNLP-08 twelve-NE tags. This conversion is a semi-automatic process. Some of our sixteen NE tags can be automatically mapped to some of the IJCNLP-08 shared task tags. The tags that represent person, location and organization names can be directly mapped to the NEP, NEL and NEO tags, respectively. Other IJCNLP-08 shared task tags can be obtained with the help of gazetteer lists and simple heuristics. In our earlier NER experiments, we have already developed a number of gazetteer lists such as: lists of person, location and organization names; list of prefix words (e.g., *sree*, *sriman* etc.) that predict the left boundary of a person name; list of designation words (e.g., *mantri*, *sangsad* etc.) that helps to identify person names. The lists of prefix and designation words are helpful to find the NETP and NED tags. The sixteen-NE tagged corpus is searched for the person name tags and the previous word is matched against the lists of prefix and designation words. The previous word is tagged as NED or NETP if there is a match with the lists of designation words and prefix words, respectively. The NEN and NETI tags can be obtained by looking at our miscellaneous tags and using some simple heuristics. The NEN tags can be simply obtained by checking whether the MISC tagged element consists of digits only. The lists of cardinal and ordinal numbers have been also kept to recognize NETI tags. A list of words that denote the measurements (e.g., *kilogram*, *taka*, *dollar* etc.) has been kept in order to get the NEM tag. The lists of words, denoting the brand names, title-objects and terms, have been prepared to get the NEB, NETO and NETE tags. The NEA tags can be obtained with the help of a gazetteer list and using some simple heuristics (whether the word contains the dot and there is no space between the characters). The mapping from our NE tagset to the IJCNLP-08 NER shared task tagset is shown in Table 8.

## 3 Use of Language Resources

The NE tagged news corpus, developed in this work, has been used to develop the Named Entity Recognition (NER) systems in Bengali using pat-

tern directed shallow parsing, HMM, ME, CRF and SVM frameworks.

NE Class	Number of wordforms	Number of distinct wordforms
Person name	5, 123	3, 201
Location name	1, 675	1, 119
Organization name	168	131
Designation	231	102
Abbreviation	32	21
Brand	15	12
Title-person	79	51
Title-object	63	42
Number	324	126
Measure	54	31
Time	337	212
Terms	46	29

Table 7. Statistics of the 30K-tagged Corpus

Sixteen-NE tagset	IJCNLP-08 tagset
PER, LOC, ORG	NEP, NEL, NEO
B-PER, I-PER, E-PER	NEP
B-LOC, I-LOC, E-LOC	NEL
B-ORG, I-ORG, E-ORG	NEO
MISC	NEN
B-MISC, I-MISC, E-MISC	NETI, NEM
Brand name gazetteer	NEB
Title-object gazetteer	NETO
Term gazetteer	NETE
Person prefix word + PER/B-PER, I-PER, E-PER	NETP
Designation word +PER/B-PER, I-PER, E-PER	NED
Abbreviation gazetteer + Heuristics	NEA

Table 8. Tagset Mapping Table

We have considered the sixteen NE tags to develop these systems. Named entity recognition in Indian Languages (ILs) in general and particularly in Bengali is difficult and challenging as there is no concept of capitalization in ILs.

The Bengali NER systems that use pattern directed shallow parsing approach can be found in

(Ekbal and Bandyopadhyay, 2007a) and (Ekbal and Bandyopadhyay, 2007b). An HMM-based Bengali NER system can be found in (Ekbal and Bandyopadhyay, 2007c). These systems have been trained and tested with the corpus tagged with the sixteen NE tags.

A number of experiments have been conducted in order to find out the best feature set for NER in Bengali using the ME, CRF and SVM frameworks. In all these experiments, we have used a number of gazetteer lists such as: first names (72, 206 entries), middle names (1,491 entries) and sur names (5,288 entries) of persons; prefix (245 entries) and designation words (947 entries) that help to detect person names; suffixes (45 and 23 entries) that help to identify person and location names; clue words (94 entries) that help to detect organization names; location name (7, 870 entries) and organization name (2,225 entries). Apart from these gazetteer lists, we have used the prefix and suffix (may not be linguistically meaningful suffix/prefix) features, digit features, first word feature and part of speech information of the words etc. We have used the C++ based Maximum Entropy package<sup>6</sup>, C++ based OpenNLP CRF++ package<sup>7</sup> and open source YamCha<sup>8</sup> tool for ME based NER, CRF based NER and SVM based NER, respectively. For SVM based NER system, we have used TinySVM<sup>9</sup> classifier, pair wise multi-class decision method and the second-degree polynomial kernel function. The brief descriptions of all the models are given below:

- A: Pattern directed shallow parsing approach without linguistic knowledge.
- B: Pattern directed shallow parsing approach with linguistic knowledge.
- HMM based NER: Trigram model with additional context dependency, NE suffix lists for handling unknown words.
- ME based NER: Contextual window of size three (current, previous and the next word), prefix and suffix of length upto three of the current word, POS information of the current word, NE information of the previous word (dynamic feature), different digit features and the various gazetteer lists.

<sup>6</sup><http://homepages.inf.ed.ac.uk/s0450736/software/maxent/maxent-20061005.tar.bz2>

<sup>7</sup><http://crfpp.sourceforge.net>

<sup>8</sup><http://chasen.org/~taku/software/yamcha/>

<sup>9</sup><http://cl.aist-nara.ac.jp/taku-ku/software/TinySVM>

•CRF based NER: Contextual window of size five (current, previous two words and the next two words), prefix and suffix of length upto three of the current word, POS information of window three (current word, previous word and the next word), NE information of the previous word (dynamic feature), different digit features and the various gazetteer lists.

•SVM based NER: Contextual window of size six (current, previous three words and the next two words), prefix and suffix of length upto three of the current word, POS information of window three (current word, previous word and the next word), NE information of the previous two words (dynamic feature), different digit features and the various gazetteer lists.

Evaluation results of the 10-fold cross validation test for all the models are presented in Table 9. Evaluation results clearly show that the SVM based NER model outperforms other models due to its efficiency to handle the non-independent, diverse and overlapping features of Bengali language.

Model	F-Score (in %)
A	74.5
B	77.9
HMM	84.5
ME	87.4
CRF	90.7
SVM	91.8

Table 9. Results of 10-fold Cross Validation Test

## 4 Conclusion

In this work we have developed a Bengali news corpus of approximately 34 million wordforms from the web archive of a leading Bengali newspaper. The *date*, *location*, *reporter* and *agency tags* present in the web pages have been automatically NE tagged. Around 150K wordforms of this partially NE tagged corpus has been manually annotated with the sixteen NE tags. We have also tagged around 30K wordforms with the twelve NE tags, defined for the IJCNLP-08 NER shared task. A tag conversion routine has also been developed in order to convert any sixteen-NE tagged corpus to the twelve-NE tagged corpus. The sixteen-NE tagged corpus of 150K wordforms has been used to

develop the NER systems using various machine-learning approaches.

This NE tagged corpus can be used for other NLP research activities such as machine translation, information retrieval, cross-lingual event tracking, automatic summarization etc.

## References

- Bertagna, M. and A. Lenci, M. Monachini and N. Calzolari. 2004. Content Interoperability of Lexical Resources, Open Issues and "MILE" Perspectives, In *Proceedings of the LREC*, 131-134.
- Bharthi, A., D.M. Sharma, V. Chaitnya, A. P. Kulkarni and R. Sanghal. 2001. LERIL: Collaborative Effort for Creating Lexical Resources. In *Proceedings of the 6<sup>th</sup> NLP Pacific Rim Symposium Post-Conference Workshop*, Japan.
- Bikel, D. M., Schwartz, R., Weischedel, R. M. 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34, 211-231.
- Calzolari, N., F. Bertagna, A. Lenci and M. Monachini. 2003. Standards and best Practice for Multilingual Computational Lexicons, MILE (the multilingual ISLE lexical entry). *ISLE Deliverable D2.2 & 3.2*.
- Chieu, H. L., Tou Ng, H. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information, In *Proc. of the 6<sup>th</sup> Workshop on Very Large Corpora*.
- De Sitter, A., Daelemans W. 2003. Information Extraction via Double Classification. In *Proceedings of International Workshop on Adaptive Text Extraction and Mining*, Dubronik.
- Ekbal, Asif, and S. Bandyopadhyay. 2007a. Pattern Based Bootstrapping Method for Named Entity Recognition. In *Proceedings of the 6<sup>th</sup> International Conference on Advances in Pattern Recognition*, 2007, India, 349-355.
- Ekbal, Asif, and S. Bandyopadhyay. 2007b. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In *Proceedings of the 5<sup>th</sup> International Conference on Natural Language Processing (ICON)*, India, 123-128.
- Ekbal, Asif, Naskar, Sudip and S. Bandyopadhyay. 2007c. Named Entity Recognition and Transliteration in Bengali, *Named Entities: Recognition, Classification and Use, Special Issue of Lingvisticae Investigationes Journal*, 30:1 (2007), 95-114.
- Ekbal, Asif, and S. Bandyopadhyay. 2007d. A Web-based Bengali News Corpus for Named Entity Recognition. *Language Resources and Evaluation Journal* (Accepted and to appear by December 2007).
- Fletcher, W. H. 2001. Making the Web More Useful as Source for Linguistics Corpora. In Ulla Conon and Thomas A. Upton (eds.), *Applied corpus Linguistics: A Multidimensional Perspective*, 191-205.
- Fletcher, W. H. 2003. Concorde the Web with KwiCFinder. In *Proceedings of the Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston.
- Giguet, E., and P. Luquet. 2006. Multilingual Lexical Database Generation from Parallel Texts in 20 European Languages with Endogeneous Resources. In *Proceedings of the COLING/ACL*, Sydney, 271-278.
- Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In *Proceedings of the 18<sup>th</sup> International Conference on Machine Learning*, 282-289.
- Lenci, A., N. Bel, F. Busu, N. Calzolari, E. Gola, M. Monachini, A. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas and A. Zampolli. 2000. SIMPLE: A general Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography, Special Issue, Dictionaries, Thesauri and Lexical-Semantic Relations*, XIII(4): 249-263.
- Li, Wei and Andrew McCallum. 2004. Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Inductions. *ACM TALIP*, Vol. 2(3), 290-294.
- McCallum, A., Freitag, D., Pereira, F. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the 17<sup>th</sup> International Conference Machine Learning*.
- Robb, T. 2003. Google as a Corpus Tool? *ETJ Journal*, 4(1), Spring 2003.
- Rundell, M. 2000. The Biggest Corpus of All. *Humanising Language Teaching*, 2(3).
- Sun, A., et al. 2003. Using Support Vector Machine for Terrorism Information Extraction. In *Proceedings of 1<sup>st</sup> NSF/NIJ Symposium on Intelligence and Security*.
- Takenobou, T., V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Soria, C. Huang, X. YingJu, Y. Hao, L. Prevot and S. Kiyooki. 2006. Infrastructure for Standardization of Asian Languages Resources. In *Proceedings of the COLING/ACL 2006*, Sydney, 827-834.



# Gazetteer Preparation for Named Entity Recognition in Indian Languages

**Sujan Kumar Saha**

Indian Institute of Technology  
Kharagpur, West Bengal  
India - 721302

sujan.kr.saha@gmail.com

**Sudeshna Sarkar**

Indian Institute of Technology  
Kharagpur, West Bengal  
India - 721302

shudeshna@gmail.com

**Pabitra Mitra**

Indian Institute of Technology  
Kharagpur, West Bengal  
India - 721302

pabitra@gmail.com

## Abstract

This paper describes our approaches for the preparation of gazetteers for named entity recognition (NER) in Indian languages. We have described two methodologies for the preparation of gazetteers<sup>1</sup>. Since the relevant gazetteer lists are more easily available in English we have used a transliteration based approach to convert available English name lists to Indian languages. The second approach is a context pattern induction based domain specific gazetteer preparation. This approach uses a domain specific raw corpus and a few seed entities to learn context patterns and then the corresponding name lists are generated by using bootstrapping.

## 1 Introduction

Named entity recognition involves locating and classifying the names in text. NER is an important task, having applications in information extraction (IE), question answering (QA), machine translation and in most other NLP applications.

NER systems have been developed for resource-rich languages like English with very high accuracies. But constructing an NER for a resource-poor language is very challenging due to unavailability of proper resources. Name-dictionaries or gazetteers are very helpful NER resources and in most Indian

<sup>1</sup>Specialized list of names for a particular class of Named Entity (NE). For Example, *India* is in the location gazetteer, *Sachin* is in the person first name gazetteer.

languages there is no reasonable size publicly available list. The web contains lots of such resources, which can be used for Indian language NER development. But most of the web resources are in English. Our approach is to transliterate the relevant English resources and name dictionaries into Indian languages to make them useful for Indian language NER task. But direct transliteration from English to an Indian languages is not easy. Few attempts are taken to build English to Indian language transliteration systems but the word agreement ratio (WAR) reached is upto 69.3% (Ekbal et al., 2006).

We have attempted to build a transliteration system which uses an intermediate alphabet. Both the English and the Indian language strings are transliterated to the intermediate alphabet and for a English-Indian language pair, if the transliterated intermediate alphabet strings are same then we have concluded that the strings are the transliteration of one another. We have transliterated the available English name lists into the intermediate alphabet and these might be used as gazetteers. The Indian language words need to be transliterated to the intermediate format to check whether the word is in a gazetteer or not. This system does not transliterate the English name lists into Indian languages but makes them useful in Indian languages NER task.

Transliteration based approaches are useful when there is availability of English name lists. But when relevant English name lists are not available then also we can prepare gazetteers from raw corpus. We have defined a semi-automatic context pattern (CP) extraction based gazetteer preparation framework. This framework uses bootstrapping to prepare

the gazetteers from a large raw corpus starting from few seed entities. Firstly fixed length patterns are formed using the surrounding words of the seeds. Depending on the pattern precision, the patterns are discarded or generalized by dropping tokens from the patterns. This set of high precision patterns extracts other named entities (NEs) which are added to the seed list for the next iteration of the process. Finally we are able to prepare the required gazetteers. To prove the effectiveness of the gazetteer preparation approach, we have prepared some gazetteers like names of cricketers, names of tennis players etc. from a raw Hindi sports domain corpus. The details of the approaches are given in the following sections.

The paper is organized as follows. Usefulness of gazetteers in NER, transliteration approaches in general and specific for Indian languages and general pattern extraction methodologies are discussed in section 2. Section 3 presents the architecture of the 2-phase transliteration system and preparation of gazetteers using that. In section 4 context pattern extraction based gazetteer preparation is discussed. Finally section 5 concludes the paper.

## 2 Previous Work

The main approaches to NER are Linguistic approaches and Machine Learning (ML) based approaches. The linguistic approach typically uses rule-based models manually written by linguists. ML based techniques make use of a large amount of annotated training data to acquire high-level language knowledge. Several ML techniques like Hidden Markov Model (HMM)(Bikel et al., 1997), Maximum Entropy Model(MaxEnt) (Borthwick, 1999), Conditional Random Field(CRF) (Li and McCallum, 2004) etc. have been successfully used for the NER task. Both the approaches may make use of gazetteer information to build systems. There are many systems which use gazetteers to improve the accuracy.

Ralph Grishman has developed a rule-based NER system which uses some specialized name dictionaries including names of all countries, names of major cities, names of companies, common first names etc (Grishman, 1995). Another rule based NER system is developed by Wakao et al. (1996) which has used

several gazetteers like organization names, location names, person names, human titles etc.

We will now mention some ML based systems. *MENE* is a MaxEnt based system developed by Borthwick. This system has used 8 dictionaries (Borthwick, 1999), which are: First names (*1,245*), Corporate names (*10,300*), Corporate names without suffix (*10,300*), Colleges and Universities (*1,225*), Corporate suffixes (*244*), Date and Time (*51*) etc. The italics numbers in bracket indicates the size of the dictionaries. The hybrid system developed by Srihari et al.(2000) combines several modules built by using MaxEnt, HMM and handcrafted rules. This system uses the following gazetteers: First name (*8,000*), Family name (*14,000*) and a big gazetteer of Locations (*250,000*). There are many other systems which have used name dictionaries to improve the accuracy. Kozareva (2006) described a methodology to generate gazetteer lists automatically for Spanish and to build NER system with labeled and unlabeled data. The location gazetteer is built by finding location patterns which looks for specific prepositions. And the person gazetteer is constructed with graph exploration algorithm.

Transliteration is also a very important topic and lots of transliteration systems for different languages have been developed using different approaches. The basic approaches for transliteration are phoneme based or spelling-based. A phoneme-based statistical transliteration system from Arabic to English was developed by Knight and Graehl(1998). This system uses a finite state transducer that implements transformation rules to do back-transliteration. A spelling-based model that directly maps English letter sequences into Arabic letters was developed by Al-Onaizan and Knight(2002). Several transliteration systems exist for English-Japanese, English-Chinese, English-Spanish and many other languages to English. But very few attempts have been reported on the development of transliteration systems between Indian languages and English. We can mention a transliteration system for Bengali-English transliteration developed by Ekbal et al.(2006). They have proposed different models modifying the joint source channel model. In that system a Bengali string is divided into transliteration units containing a vowel

modifier or *matra* at the end of each unit. Similarly the English string is also divided into units. Then they defined various unigram, bigram or trigram models depending on the consideration of the contexts of the units. They have also considered linguistic knowledge in the form of possible conjuncts and diphthongs in Bengali and their representations in English. This system is capable of transliterating mainly person names. The highest transliteration accuracy achieved by them is 69.3% Word Agreement Ratio (WAR) for Bengali to English and 67.9% WAR for English to Bengali transliteration.

In the field of IE, patterns play a key role in identifying relevant pieces of information. Soderland et al.(1995), Rilof and Jones(1999), Lin et al.(2003), Downey et al.(2004), Etzioni et al.(2005) described different approaches to context pattern induction. Talukder et al.(2006) combined grammatical and statistical techniques to create high precision patterns specific for NE extraction. An approach to lexical pattern learning for Indian languages is described by Ekbal and Bandopadhyay (2007). They used seed data and annotated corpus to find the patterns for NER.

### 3 Transliteration based Gazetteer Preparation

Gazetteers or name dictionaries are helpful in NER. We have already discussed about some English NER systems where the usefulness of the gazetteers have been established. However while developing NER systems in Indian languages, we tried to find relevant gazetteers. But we could not obtain openly available gazetteer lists for these languages. But we found that there are a lot of resources of names of Indian persons, Indian places, organizations etc. in English available in the web. In Table 1 we have mentioned some of the sources which contains relevant name lists.

But it is not possible to use the available name lists directly in the Indian language NER task as these are in English. We have decided to transliterate the English lists into Indian languages to make them useful in the Indian language NER task.

List	Web Sources
First Name	<a href="http://www.bsnl.co.in/onlinedirectory.htm">http://www.bsnl.co.in/ onlinedirectory.htm</a> <a href="http://web1.mtnl.net.in/directory/">http://web1.mtnl.net.in/ directory/</a> <a href="http://www.eci.gov.in/">http://www.eci.gov.in/</a> <a href="http://hiren.info/indian-baby-names/">http://hiren.info/indian-baby-names/</a> <a href="http://www.indiaexpress.com/specials/babynames/">http://www.indiaexpress.com/specials/babynames/</a>
Surname	<a href="http://surnamedirectory.com/surname-index.html">http://surnamedirectory.com/surname-index.html</a> <a href="http://web1.mtnl.net.in/directory/">http://web1.mtnl.net.in/ directory/</a> <a href="http://en.wikipedia.org">http://en.wikipedia.org</a>
India Location	<a href="http://indiavilas.com/indainfo/pincodes.asp">http://indiavilas.com/ indainfo/pincodes.asp</a> <a href="http://www.indiapost.gov.in">http://www.indiapost.gov.in</a> <a href="http://www.eci.gov.in/">http://www.eci.gov.in/</a>
World Location	<a href="http://www.maxmind.com/app/worldcities">http://www.maxmind.com/app/worldcities</a> <a href="http://en.wikipedia.org/wiki">http://en.wikipedia.org/wiki</a>

Table 1: Web sources for some relevant name lists

#### 3.1 Transliteration

The transliteration from English to Hindi is quite difficult. English alphabet contains 26 characters whereas the Hindi alphabet contains 52 characters. So the mapping is not trivial. We have already mentioned that for Bengali a transliteration system was developed by Ekbal et al. Similar approach can be used to develop transliteration systems for other Indian languages. But this approach uses a bilingual transliteration corpus, which requires much efforts to built, is unavailable in proper size in all Indian languages. Also using this approach the word agreement ratio obtained is below 70%.

To make the transliteration process easier and more accurate, we have decided to build a 2-phase transliteration module. Our goal is to make decision that a particular Indian language string is in an English gazetteer or not. We need not transliterate directly from Indian language strings to English or English name lists into Indian languages. Our idea is to define an intermediate alphabet and both English and Indian language strings will be transliterated to

the intermediate alphabet. For two English-Hindi string pair, if the intermediate alphabet is same then we can conclude that one string is the transliteration of the other.

First of all we need to decide the size of the intermediate alphabet. Preserving the phonetic properties we have defined our intermediate alphabet consisting of 34 characters. To indicate these 34 characters, we have given unique character-id to each character.

### 3.2 English to Intermediate Alphabet Transliteration

For transliterating English strings into the intermediate state, we have built a phonetic map table. This phonetic map table maps an English n-gram into an intermediate character. A part of the map table is given in Table 2. In the map table, the mapping is from strings of varying length in the English to one character in the intermediate alphabet. In our table the length of the left hand side varies from 1 to 3.

English	Intermediate
A	â
EE, I, II	î
OO, U	û
B, W	ô
BH, V	ô
CH	ç
R, RH	ř
SH, S	š

Table 2: A part of the map table

In the following we have described the procedure of transliteration.

#### Procedure 1: Transliteration

Source string - English, Output string - Intermediate.

1. Scan the source string (S) from left to right.
2. Extract the first n-gram (G) from the string. ( $n = 3$ )
3. Find if it is in the map-table.
4. If yes, insert its corresponding intermediate state entity into target string T.  
Remove the n-gram from S.

$$S = S - G.$$

Go to step 2.

5. Else set  $n = n - 1$ .

Go to step 3.

Here we can take an Indian language name, ‘surabhi’, as example to explain the procedure in details. When the name is written in English, it may be written in several ways like ‘suravi’, ‘shuravi’, ‘surabhee’, ‘shurabhi’ etc. The English string ‘surabhi’ is transliterated to ‘šûřâvî’ by the transliterator. Again if we see the transliteration for ‘shuravi’, then also the intermediate transliterated string is same as the previous one.

### 3.3 Indian Language to Intermediate Alphabet Transliteration

This is a 2-phase process. The first phase transliterates the Indian language string into itrans. Itrans is representation of Indian language alphabets in terms of ASCII. Since Indian text is composed of syllabic units rather than individual alphabetic letters, itrans uses combinations of two or more letters of English alphabet to represent an Indian language syllable. However, there being multiple sounds in Indian languages corresponding to the same English letter, not all Indian syllables can be represented by logical combinations of English alphabet. Hence, itrans uses some non-alphabetic special characters also in some of the syllables. A map table<sup>2</sup>, with some heuristic knowledge, is used for the transliteration. For example, the Hindi word ‘surabhi’ is converted ‘sUrabhI’ in itrans.

In the second phase the itrans string is transliterated into the intermediate state using the similar procedure described section 3.2. Here also we use a map-table containing the mappings from itrans to intermediate alphabet. This procedure transliterates the example itrans word ‘sUrabhI’ to ‘šûřâvî’.

### 3.4 Evaluation

In section 3.2 and 3.3 we have described two phase transliteration with an example word. We have shown that our transliteration system transliterates the Indian language name ‘surabhi’ and the corresponding English strings into the same intermediate

<sup>2</sup>The map table is available at [www.aczoom.com/itrans](http://www.aczoom.com/itrans).

string. The system has limitations like sometimes two different strings can be mapped into a same intermediate alphabet string.

For the evaluation of the system we have applied the transliteration system for two languages - Hindi and Bengali. For evaluating the system for Hindi we have created a bi-lingual corpus containing 1070 English-Hindi word pair most of which are names. 980 of them are transliterated correctly by the system. The system accuracy is  $980 \times 100/1070 = 91.59\%$ . For evaluating the system for Bengali, we have used a similar bi-lingual corpus and the system transliterates with 89.3% accuracy.

### 3.5 Prepared Gazetteer Lists

Previously we have mentioned the web sources where some name lists are available. Names of a particular category are collected from different sources and merged to build a English name list of that category. Then we have applied our transliteration procedure on the list and transliterated the list into the intermediate alphabet. This intermediate alphabet list acts as a gazetteer in NER task in Indian languages. When an Indian language NER system needs to access the gazetteer lists, it transliterates the Indian language strings into the intermediate alphabet, and searches into the list. In the following we have described the prepared gazetteer lists which are useful for a general domain Indian language NER system.

**First Name List:** This list contains 10,200 first names collected from the web. Most of the collected first names are of Indian origin. Apart from the Indian names, we have also collected some non-Indian names. These non-Indian names are generally the names of some famous persons, like sports stars, film stars, scientists, politicians, who are likely to come in Indian language texts. In our first name list 1500 such names are included.

**Surname List:** This is a very important list which contains common surnames. We have prepared the surname list from different sources containing about 1500 Indian surnames and 400 other surnames.

**Indian Locations:** This list contains about 14,000 entities. The names of states, cities and towns, districts, important places in different cities and even lots of village names are collected in the list. The list needs to be processed into a list of un-

igrams (e.g., *kolakAtA*<sup>3</sup> (Kolkata), *bihAra* (Bihar)), bigrams (e.g., *nayI dilli* (New Delhi), *pashchima bA.nglA* (West Bengal)) and trigrams (e.g. *uttaara chabisha paraganA* (North 24 Pargana)). The words are matched with unigrams, sequences of two consecutive words are matched against bigrams and sequences of three consecutive words are matched against trigrams.

**World Location:** The list contains the names of the countries, different state and city names in world and also the names of important rivers, mountains etc. The list contains about 4,000 location names. Similar to the Indian location list, this list also needs to be processed as unigram, bigrams and trigrams.

### 4 Context Pattern Extraction based Gazetteer Preparation

Gazetteers can also be prepared by extracting context patterns. Transliteration based gazetteer preparation is applicable while there is availability of English or parallel language name list. But if such relevant name lists are not available, but a large raw corpus is available, then we can use the context pattern extraction based methodology to prepare the gazetteers. This method seeks some high precision context patterns by using some seed entities and hits the patterns to the raw corpus to prepare the gazetteers.

The overall methodology of extracting context patterns from a raw corpus is summarized as follows:

1. Find a large raw corpus and some seed entities ( $E$ ) for each class of NEs.
2. For each seed entity  $e$  in  $E$ , from the corpus find context string( $C$ ) comprised of  $n$  tokens before  $e$ , a placeholder for the class instance and  $n$  tokens after  $e$ . [We have used  $n = 3$ ] This set of words form initial pattern.
3. Search the pattern to the corpus and find the coverage and precision.
4. Discard the patterns having low precision.
5. Generalize the patterns by dropping one or more tokens to increase coverage.

<sup>3</sup>The Indian languages strings are written in italics font and using itrans transliteration.

6. Find best patterns having good precision and coverage.

The details of the context pattern extraction based gazetteer preparation methodology is described in the following subsections. We have taken a Hindi sports domain raw corpus and prepared some gazetteers like names of cricketers, names of tennis players to prove the effectiveness of the proposed methodology.

#### 4.1 Selection of Seed Entity

Context pattern extraction based gazetteer preparation methodology is applied to a sports domain corpus which contains about 20 lakhs words collected from the popular Hindi newspaper “Dainik Jagaran”. We have worked on preparing the lists of cricket players, list of tennis players. We have collected the most frequent names to build the seed list. For preparing the list of tennis players, we have taken 5 names as seed entities : Andre Agassi, Steffi Graf, Serena Williams, Roger Federer and Justine Henin. Similarly the seed list of cricket players name list contains only 3 names: Sachin Tendulkar, Brian Lara and Glenn McGrath.

#### 4.2 Context Extraction

To extract the patterns for a particular category, we select a part of the corpus where the target seeds will be available with high frequency. For example to get the patterns for the names of the cricketers, we select a part of the corpus where most of the sentences are cricket related. To select the cricket related sentences, we prepared a list containing the most frequent words related to cricket like, *rAna* (run), *ballebAja* (batsman), *gedabAja* (bowler) etc. Depending upon the presence of such words we have selected the ‘part’. In our development the cricket ‘part’ contains 120K words. Similar ‘part’ is developed for other gazetteers. For a particular seed, we find the occurrences of the seed entity in the corresponding raw ‘part’ corpus. Then we extracted three tokens immediately preceding the seed and three tokens immediately following the seed. A placeholder (*CRIC* for cricketers, *TENS* for tennis players) replaces the seed. The placeholder and the surrounding tokens  $t_{-3}$   $t_{-2}$   $t_{-1}$  *placeholder*  $t_{+1}$   $t_{+2}$   $t_{+3}$ ) form the initial set of patterns.

For the seed *sachina tedulkara* (Sachin Tendulkar) we extract 92 initial patterns. Some of which are:

- *ki mAsTara blAsTara CRIC ko Tima ke*
- *dravi.Da aura CRIC 241 nAbAda ke*
- *bhAratiya ballebAja CRIC ne 100 rana*
- *mere vichAra se CRIC ko Takkara dene*

#### 4.3 Pattern Quality Measure

We measured the quality of a pattern depending on its precision and coverage. Precision is the ratio of correct identification and the total identification. If the precision is high then also we have assumed that the pattern is a *good* pattern. In our development we have marked a pattern as *good* if the precision is 100%.

We search the initial patterns in the corresponding ‘part’ corpus to measure the precision and coverage. If the precision is less than 100% for a pattern then we have rejected the pattern. Otherwise we have tried to make it more generalized to increase the coverage. To make the generalization we have dropped the left most and right most tokens one by one and checked the pattern quality. If for a initial pattern, several patterns presents with 100% precision then we have selected those patterns for which no subset of those is a *good* pattern. By this way we have prepared a list of *good* patterns for a particular gazetteer type.

In time of ‘good’ pattern selection we have made some interesting observations.

- There are some patterns which satisfy the 100% precision criteria but the coverage is very poor in terms of new entity extraction. For example, “*mAsTara blAsTara CRIC ko*” is a pattern with 100% precision. The pattern has 24 instances in the ‘part’ corpus, but all the extracted entities are ‘*sachina tedulkara*’. We have also examined the pattern in the total raw corpus. It is capable of extracting ‘*sachina tedulkara*’ only. So in spite of fulfilling all the criteria the pattern is not a ‘good’ pattern.
- Another interesting observation is, that there are some patterns which are ‘good’ patterns in

the context of the ‘part’ corpus, but when used in the total raw corpus, it extracts non-relevant entities. For example “mere vichAra se *CRIC* ko Takkara” is a ‘good’ pattern so it should extract the names of the cricketers. But when this is used in the total raw corpus it extracts non-cricketer entities (e.g. tennis players, chess players) also. To make such patterns useful we have extracted all the cricket related sentences in the similar way which was used for selecting the ‘part’ corpus and then the patterns are used to extract entities from these sentences.

- There present are patterns with very high coverage but precision is just below 100%. We have analyzed these patterns and manually identified the wrongly extracted entities. If the wrong entities can be grouped together and are having some specific properties then we have added these entities in a ‘pattern exception list’. Then the pattern is used as a ‘good’ pattern and the exception list is used to detect the wrong identifications.

In the following we have given some example of ‘good’ patterns which are useful in identification of the names of the cricket players.

- ballebAja *CRIC* ko Tima ke
- ballebAja *CRIC* ne
- *CRIC* kA arddhashataka

#### 4.4 Gazetteer Preparation

The extracted ‘good’ patterns are capable of identifying NEs from a raw corpus. These patterns are then used to prepare the gazetteers. The seeds form the initial gazetteer list for a particular gazetteer type. The ‘good’ patterns are used to extract entities from the total raw corpus. The entities identified by the patterns are added to the corresponding gazetteer list. In that way we can add more entities in our first phase gazetteer list. These new entities are taken as seeds for the next phase. Then the same procedure is followed repeatedly to develop a large gazetteer.

We have already mentioned that we have worked with a sports domain corpus and prepared some gazetteers. This gazetteers are prepared just to prove the efficiency of our approach. By using only 3 seed

entities we become able to prepare a gazetteer which contains 412 names of the cricketers. Even using this approach only one seed ‘Sachin Tendulkar’ extracts 297 names after the second iteration. Similarly we have collected 245 names of tennis players from 5 seed entities.

## 5 Conclusion

In this paper we have described our approaches for the preparation of gazetteers. We have also prepared some gazetteers using both the approaches to show their effectiveness. These approaches are very useful for the NER task in resource-poor languages and also in domain specific NER task.

## References

- Al-Onaizan Y. and Knight K. 2002. Machine Transliteration of Names in Arabic Text. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*.
- Bikel Daniel M., Miller Scott, Schwartz Richard and Weischedel Ralph. 1997. Nymble: A high performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201.
- Borthwick Andrew. 1999. A Maximum Entropy Approach to Named Entity Recognition. *Ph.D. thesis, Computer Science Department, New York University*.
- Downey D., Etzioni O., Soderland S., Weld D.S. 2004. Learning text patterns for Web information extraction and assessment. In *AAAI-04 Workshop on Adaptive Text Extraction and Mining*, pages 50–55.
- Ekbal A. and Bandyopadhyay S. 2007. Lexical Pattern Learning from Corpus Data for Named Entity Recognition. In *Proceedings of International Conference on Natural Language Processing (ICON), 2007*.
- Ekbal A., Naskar S. and Bandyopadhyay S. 2006. A Modified Joint Source Channel Model for Transliteration. In *Proceedings of the COLING/ACL 2006, Australia*, pages 191–198.
- Etzioni Oren, Cafarella Michael, Downey Doug, Popescu Ana-Maria, Shaked Tal, Soderland Stephen, Weld Daniel S. and Yates Alexander. 2005. Unsupervised named-entity extraction from the Web: An experimental study. In *Artificial Intelligence*, 165(1): 91-134.
- Grishman Ralph. 1995. The New York University System MUC-6 or Where’s the syntax? In *Proceedings of the Sixth Message Understanding Conference*.

- Knight K. and Graehl J. 1998. Machine Transliteration. *Computational Linguistics*, 24(4): 599–612.
- Kozareva Zornitsa. 2006. Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists. In *Proceedings of EACL student session (EACL 2006)*.
- Li Wei and McCallum Andrew. 2004. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction (Short Paper). In *ACM Transactions on Computational Logic*.
- Lin Winston, Yangarber Roman and Grishman Ralph. 2003. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*.
- Riloff E. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049.
- Srihari R., Niu C. and Li W. 2000. A Hybrid Approach for Named Entity and Sub-Type Tagging. In *Proceedings of the sixth conference on Applied natural language processing*.
- Soderland Stephen, Fisher David, Aseltine Jonathan, Lehnert Wendy. 1995. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*.
- Talukdar P. Pratim, T. Brants, M. Liberman and F. Pereira. 2006. A context pattern induction method for named entity extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*.
- Wakao T., Gaizauskas R. and Wilks Y. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of COLING-96*.



# Preliminary Chinese Term Classification for Ontology Construction

Gaoying Cui, Qin Lu, Wenjie Li

Department of Computing,

Hong Kong Polytechnic University

{csgycui, csluqin, cswjli}@comp.polyu.edu.hk

## Abstract

An ontology can be seen as a representation of concepts in a specific domain. Accordingly, ontology construction can be regarded as the process of organizing these concepts. If the terms which are used to label the concepts are classified before building an ontology, the work of ontology construction can proceed much more easily. Part-of-speech (PoS) tags usually carry some linguistic information of terms, so PoS tagging can be seen as a kind of preliminary classification to help constructing concept nodes in ontology because features or attributes related to concepts of different PoS types may be different. This paper presents a simple approach to tag domain terms for the convenience of ontology construction, referred to as *Term PoS (TPoS) Tagging*. The proposed approach makes use of segmentation and tagging results from a general PoS tagging software to predict tags for extracted domain specific terms. This approach needs no training and no context information. The experimental results show that the proposed approach achieves a precision of 95.41% for extracted terms and can be easily applied to different domains. Comparing with some existing approaches, our approach shows that for some specific tasks, simple method can obtain very good performance and is thus a better choice.

**Keywords:** ontology construction, part-of-speech (PoS) tagging, Term PoS (TPoS) tagging.

## 1 Introduction

Ontology construction has two main issues including the acquisition of domain concepts and the acquisition of appropriate taxonomies of these concepts. These concepts are labeled by the terms used in the domain which are described by different attributes. Since domain specific terms (terminology) are labels of concepts among other things, terminology extraction is the first and the foremost important step of domain concept acquisition. Most of the existing algorithms in Chinese terminology extraction only produce a list of terms without much linguistic information or classification information (Yun Li and Qiangjun Wang, 2001; Yan He et al., 2006; Feng Zhang et al., 2006). This fact makes it difficult in ontology construction as the fundamental features of these terms are missing. The acquisition of taxonomies is in fact the process of organizing domain specific concepts. These concepts in an ontology should be defined using a subclass hierarchy by assigning and defining properties and by defining relationship between concepts etc. (Van Rees, R., 2003). These methods are all concept descriptions. The linguistic information associated with domain terms such as PoS tags and semantic classification information of terms can also make up for the concept related features which are associated with concept labels. Terms with different PoS tags usually carry different semantic information. For example, a noun is usually a word naming a thing or an object. A verb is usually a word denoting an action, occurrence or state of existence, which are all associated with a time and a place. Thus in ontology construction, noun nodes and verb nodes should be described using different attributes with different discriminating characters. With this information, extracted terms can then be classified

accordingly to help in ontology construction and retrieval work. Thus PoS tags can help identify the different features needed for concept representation in domain ontology construction.

It should be pointed out that *Term PoS (TPoS)* tagging is different from the general PoS tagging tasks. It is designed to do PoS tagging for a given list of terms extracted from some terminology extraction algorithms such as those presented in (Luning Ji et al., 2007). The granularity of general PoS tagging is smaller than what is targeted in this paper because terms representing domain specific concepts are more likely to be compound words and sometimes even phrases, such as “文件管理器”(file manager), “并发描述”(description of concurrency), etc.. Even though current general word segmentation and PoS tagging can achieve precision of 99.6% and 97.58%, respectively (Huaping Zhang et al., 2003), its performance for domain specific corpus is much less satisfactory (Luning Ji et al., 2007), which is why terminology extraction algorithms need to be developed.

In this paper, a very simple but effective method is proposed for TPoS tagging which needs no training process or even context information. This method is based on the assumption that every term has a headword. For a given list of domain specific terms which are segmented and each word in the term already has a PoS tag, the TPoS tagging algorithm then identifies the position of the headword and take the tag of the headword as the tag of the term. Experiments show that this method is quite effective in giving good precision and minimal computing time.

The remaining of this paper is organized as follows. Section 2 reviews the related work. Section 3 gives the observations to the task and corresponding corpus, then presents our method for TPoS tagging. Section 4 gives the evaluation details and discussions on the proposed method and reference methods. Section 5 concludes this paper.

## 2 Related Work

Although TPoS tagging is different from general PoS tagging, the general POS tagging methods are worthy of referencing. There are a lot of existing POS tagging researches which can be classified into following categories in general. Natural ideas of solving this problem were to make use of the

information from the words themselves. A number of features based on prefixes and suffixes and spelling cues like capitalization were adopted in these researches (Mikheev, A, 1997; Brants, Thorsten, 2000; Mikheev, A, 1996). Mikheev presented a technique for automatically acquiring rules to guess possible POS tags for unknown words using their starting and ending segments (Mikheev, A, 1997). Through an unsupervised process of rule acquisition, three complementary sets of word-guessing rules would be induced from a general purpose lexicon and a raw corpus: prefix morphological rules, suffix morphological rules and ending-guessing rules (Mikheev, A, 1996). Brants used the linear interpolation of fixed length suffix model for word handling in his POS tagger, named TnT. For example, an English word ending in the suffix *-able* was very likely to be an adjective (Brants, Thorsten, 2000).

Some existing methods are based on the analysis of word morphology. They exploited more features besides morphology or took morphology as supplementary means (Toutanova et al., 2003; Huihsin Tseng et al., 2005; Samuelsson, Christer, 1993). Toutanova et al. demonstrated the use of both preceding and following tag contexts via a dependency network representation and made use of some additional features such as lexical features including jointly conditioning on multiple consecutive words and other fine-grained modeling of word features (Toutanova et al., 2003). Huihsin et al. proposed a variety of morphological word features, such as the tag sequence features from both left and right side of the current word for POS tagging and implemented them in a Maximum Entropy Markov model (Huihsin Tseng et al., 2005). Samuelsson used n-grams of letter sequences ending and starting each word as word features. The main goal of using Bayesian inference was to investigate the influence of various information sources, and ways of combining them, on the ability to assign lexical categories to words. The Bayesian inference was used to find the tag assignment  $T$  with highest probability  $P(TM, S)$  given morphology  $M$  (word form) and syntactic context  $S$  (neighboring tags) (Samuelsson, Christer, 1993).

Other researchers inclined to regard this POS tagging work as a multi-class classification problem. Many methods used in machine learning, such

as Decision Tree, Support Vector Machines (SVM) and  $k$ -Nearest-Neighbors ( $k$ -NN), were used for guessing possible POS tags of words (G. Orphanos and D. Christodoulakis, 1999; Nakagawa T, 2001; Maosong Sun et al., 2000). Orphanos and Christodoulakis presented a POS tagger for Modern Greek and focused on a data-driven approach for the induction of decision trees used as disambiguation or guessing devices (G. Orphanos and D. Christodoulakis, 1999). The system was based on a high-coverage lexicon and a tagged corpus capable of showing off the behavior of all POS ambiguity schemes and characteristics of words. Support Vector Machine is a widely used (or effective) classification approach for solving two-class pattern recognition problems. Selecting appropriate features and training effective classifiers are the main points of SVM method. Nakagawa et al. used substrings and surrounding context as features and achieve high accuracy in POS tag prediction (Nakagawa T, 2001). Furthermore, Sun et al presented a POS identification algorithm based on  $k$ -nearest-neighbors ( $k$ -NN) strategy for Chinese word POS tagging. With the auxiliary information such as existing tagged lexicon, the algorithm can find out  $k$  nearest words which were mostly similar with the word need tagging (Maosong Sun et al., 2000).

### 3 Algorithm Design

As pointed out earlier, TPoS tagging is different from the general PoS tagging tasks. In this paper, it is assumed that a terminology extraction algorithm has already obtained the PoS tags of individual words. For example, in the segmented and tagged sentence “计算机/n 图形/n 学/v 中/f , /w 物体/n 常常/d 用/v 多边形/a 网格/n 来/f 表示/v 。 /w”(In computer graphics, objects are usually represented as polygonal meshes.), the term “多边形网格” (polygonal meshes) has been segmented into two individual words and tagged as “多边形/a” (polygonal /a) and “网格/n” (meshes /n). The terminology extraction algorithm would identify these two words “多边形/a” and “网格/n” as a single term in a specific domain. The proposed algorithm is to determine the PoS of this single term “多边形网格” (polygonal meshes), thus the algorithm is referred to as TPoS tagging. It can be seen that the general purpose PoS tagging and term PoS tagging assign tags at different granularity. In

principle, the context information of terms can help TPoS tagging and the individual PoS tags may be good choices as classification features.

The proposed TPoS tagging algorithm consists of two modules. The first module is a terminology extraction preprocessing module. The second module carries out the TPoS tag assignment. In the terminology extraction module, if the result of terminology extraction algorithm is a list of terms without PoS tags, a general purpose segmenter called ICTCLAS<sup>1</sup> will be used to give PoS tags to all individual words. ICTCLAS is developed by Chinese Academy of Science, the precision of which is 97.58% on tagging general words (Huaping Zhang et al., 2003). Then the output of this module is a list of terms, referred to as TermList, using algorithms such as the method described in (Luning Ji et al., 2007).

In this paper, two simple schemes for the term PoS tag assignment module are proposed. The first scheme is called the *blind assignment scheme*. It simply assigns the noun tag to every term in the TermList. This is based on the assumption that most of the terms in a specific domain represent certain concepts that are most likely to be nouns. Result from this blind assignment scheme can be considered as the baseline or the worse case scenario. Even in general domain, it is observed that nouns are in the majority of Chinese words with more than 50% among all different PoS tags (Hui Wang, 2006).

The second scheme is called *head-word-driven assignment scheme*. Theoretically, it will take the tag of the head word of one term as the tag of the whole term. But here it simply takes the tag of the last word in a term. This is based on the assumption that each term has a headword which in most cases is the last word in a term (Hui Wang, 2006). One additional experiment has been done to verify this assumption. A manually annotated Chinese shallow Treebank in general domain is used for the statistic work (Ruifeng Xu et al., 2005). There are 9 different structures of Chinese phrases, (Yunfang Wu et al., 2003), but only 3 of them do not have their head words in the tail, which are about 6.56% from all phrases. Following the examples earlier,

---

<sup>1</sup> Copyright © Institute of Computing Technology, Chinese Academy of Sciences

the term “多边形/a 网格/n” (polygonal /a meshes /n) will be assigned the tag “/n” because the last word is labeled “/n”.

There are a lot of semanteme tags at the end of a term. For example, “/ng” presents single character postfix of a noun. But it would be improper if a term is tagged as “/ng”. For example, the term “决策器” (decision-making machine) contains two segments as listed with two components “决策/n” and “器/ng”. It is obvious that “决策器/ng” is inappropriate. Thus the head-word-driven assignment scheme also includes some rules to correct this kind of problems. As will be discussed in the experiment, the current result of TPoS tagging is based on 2 simple induction rules applied in this algorithm.

#### 4 Experiments and Discussions

The domain corpus used in this work contains 16 papers selected from different Chinese IT journals between 1998 and 2000 with over 1,500,000 numbers of characters. They cover topics in IT, such as electronics, software engineering, telecom, and wireless communication. The same corpus is used by the terminology extraction algorithm developed in (Luning Ji et al., 2007). In the domain of IT, two TermLists are used for the experiment. TermList1 is a manually collected and verified term list from the selected corpus containing a total of 3,343 terms. TermList1 is also referred to as the standard answer set to the corpus for evaluation purposes. TermList2 is produced by running the terminology extraction algorithm in (Van Rees, R, 2003). TermList2 contains 2,660 items out of which 929 of them are verified as terminology and 1,731 items are not considered terminology according to the standard answer above.

To verify the validity of the proposed method to different domains, a term list containing 366 legal terms obtained from Google searching results for “法律术语大全”(complete dictionary of legal terms) is selected for comparison, which is named TermList3.

##### 4.1 Experiment on the Blind Assignment Scheme

The first experiment is designed to examine the proportion of nouns in TermList1 and TermList3, to validate of the assumption of the blind assign-

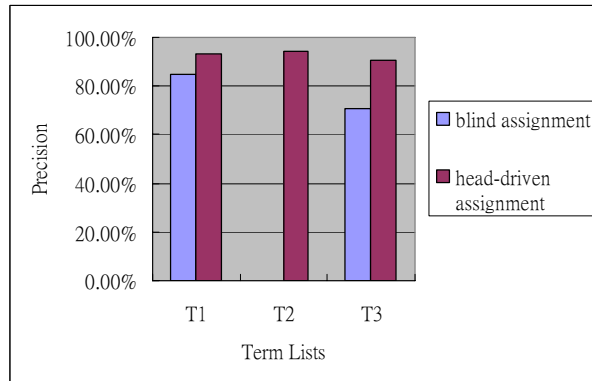
ment scheme. In first part of this experiment, all the 3,343 terms in TermList1 are tagged as nouns. The result shows that the precision of the blind assignment scheme is between 78.79% and 84.77%. The reason for the range is that there are about 200 terms in TermList1 which can be considered either as nouns, gerunds, or even verbs without reference to context. For example, the term “局域网远程访问” (“remote access of local area network” or “remote access to local area network”) and the term “极化” (polarization *or* polarize), can be considered either as nouns if they are regarded as courses of events or as verbs if they refer to the actions for completing certain work. The specific type is dependent on the context which is not provided without the use of a corpus. However, the experiment result does show that in a specific domain, there is a much higher percentage of terms that are nouns than other tags in general (Hui Wang, 2006). As to TermList3, the precision of blind assignment is between 65.57% and 70.77% (19 mixed ones). TermList2 is the result of a terminology extraction algorithm and there are non-term items in the extraction result, so the blind assignment scheme is not applied on TermList2. The blue colored bars (lighter color) in **Figure 1** shows the result of TermList1 and TermList3 using the blind assignment scheme which gives the two worst result compared to our proposed approach to be discussed in Section 4.2

##### 4.2 Experiments on the Head-Word-Driven Assignment Scheme

The experiment in this section was designed to validate the proposed head-word-driven assignment scheme. The same experiment is conducted on the three term lists respectively, as shown in **Figure 1** in purple color (darker color). The precision for assigning TPoS tags to TermList1 is 93.45%. By taking the result from a terminology extraction algorithm without regards to its potential error propagations, the precision of the head-word-driven assignment scheme for TermList2 is 94.32%. For TermList3, the precision of PoS tag assignment is 90.71%. By comparing to the blind assignment scheme, this algorithm has reasonably good performance for all three term list with precision of over 90%. It also gives 8.7% and 19.9% improvement for TermList1 and TermList3, respectively, as compared to the blind assignment

scheme, a reasonably good improvement without a heavy cost. However, there are some abnormalities in these results. Supposedly, TermList1 is a hand verified term list in the IT domain and thus its result should have less noise and thus should perform better than TermList2, which is not the case as shown in **Figure 1**.

**Figure 1** Performance of the Two Assignment Schemes on the Three Term Lists



By further analyzing the error result, for example for TermList1, among these 3,343 terms, about 219 were given improper tags, such as the term “图形学” (Graphics). In this example, two individual words, “图形/n” and “学/v”, form a term. So the output was “图形学/v” for taking the tag of the last segment. It was a wrong tag because the whole term was a noun. In fact, the error is caused by the general word PoS tagging algorithm because without context, the most likely tagging of “学”, a semanteme, is a verb. This kind of errors in semanteme tagging appeared in the results of all three term lists with 169 from TermList1, 29 from TermList2 and 12 from TermList3, respectively. This was a kind of errors which can be corrected by applying some simple induction rules. For example, for all semantemes with multiple tags (including noun as in the example), the rule can be “tagging terms with noun suffixes as nouns”. For example, terms “劳改/n 场/q” (reform-through-labor camp) and “计算机/n 图形/n 学/v” (computer graphics) were given different tags using the head-word-driven assignment scheme. They were assigned as: “劳改场/q” and “计算机图形学/v” which can be corrected by this rule. Another kind of mistake is related to the suffix tags such as “/ng” (noun suffix) and “/vg”(verb suffix). For examples, “知识/n 产权/n 庭/ng” (intellectual property tri-

bunal) and “数据/n 集/vg” (data set) will be tagged as “知识产权庭/ng” and “数据集/vg”, respectively, which are obviously wrong. So, the simple rule of “tagging terms with “/ng” and “/vg” to “/n” is applied. The performance of TPoS tag assignment after applying these two fine tuning induction rules are shown in **Table 1** below.

**Table 1** Influence of Induction Rules on Different Term Lists

Term Lists	Precision of tagging	Precision after adding induction rule	Improvement Percentage
TermList1	93.45%	97.03%	3.83%
TermList2	94.32%	95.41%	1.16%
TermList3	90.71%	93.99%	3.62%

It is obvious that with the use of fine tuning using induction rules, the results are much better. In fact the result for TermList1 reached 97.03% which is quite close to PoS tagging of general domain data. The abnormality also disappeared as the performance of TermList1 has the best result. The improvement to TermList2 (1.16%) is not as obvious as that for TermList1 and TermList3, which are 3.83% and 3.62%, respectively. This, however, is reasonable as TermList2 is produced directly from a terminology extraction algorithm using a corpus, thus, the results are noisier.

Further analysis is then conducted on the result of TermList2 to examine the influence of non-term items to this term list. The non-term items are items that are general words or items cannot be considered as terminology according to the standard answer sheet. For example, neither of the terms “问题” (problem) and “模式训练是” (pattern training is) were considered as terms because the former was a general word, and the latter should be considered as a fragment rather than a word. In fact, in 2,660 items extracted by the algorithm as terminology, only 929 of them are indeed terminology (34.92%), and rest of them do not qualify as domain specific terms. The result of this analysis is listed in **Table 2**.

**Table 2** Data Distribution Analysis on TermList2

	Without Induction Rules		Induction Rules Applied	
	correct terms	precision	correct terms	precision
Terms (929)	879	94.62%	898	96.66%
Non-terms (1,731)	1,630	94.17%	1,640	94.74%
Total (2,660)	2,509	94.32%	2,538	95.41%

Results show that 31 and 50 from the 929 correct terms were assigned improper PoS tags using the proposed algorithm with and without the inductions rules, respectively. That is, the precisions for correct data are comparable to that of TermList1 (93.45% and 97.03%, respectively). For the non-terms, 91 items and 101 items from 1,731 items were assigned improper tags with and without the induction rules, respectively. Even though the precisions for terms and non-terms without using the induction rules are quite the same (94.62% vs. 94.17%), the improvement for the non-terms using the induction rules are much less impressive than that for the terms. This is the reason for the relatively less impressed performance of induction rules for TermList2. It is interesting to know that, even though the performance of the terminology extraction algorithm is quite poor with precision of only around 35% (929 out of 2,666 terms), it does not affect too much on the performance of the TPoS proposed in this paper. This is mainly because the items extracted are still legitimate words, compounds, or phrases which are not necessarily domain specific.

The proposed algorithm in this paper use minimum resources. They need no training process and even no context information. But the performance of the proposed algorithm is still quite good and can be directly used as a preparation work for domain ontology construction because of its precision of over 95%. Other PoS tagging algorithms reach good performance in processing general words. For example, a k-nearest-neighbors strategy to identify possible PoS tags for Chinese words can reach 90.25% for general word PoS tagging (Maosong Sun et al., 2000). Another method based on SVM method on English corpus can reach 96.9% in PoS tagging known and unknown words (Nakagawa T, 2001). These results show that pro-

posed method in this paper is comparable to these general PoS tagging algorithms in magnitude. Of course, one main reason of this fact is the difference in its objectives. The proposed method is for the PoS tagging of domain specific terms which have much less ambiguity than tagging of general text. Domain specific terms are more likely to be nouns and there are some rules in the word-formation patterns while general PoS tagging algorithms usually need training process in which large manually labeled corpora would be involved. Experiment results also show that this simple method can be applied to data in different domains.

## 5 Conclusion and Future Work

In this paper, a simple but effective method for assigning PoS tags to domain specific terms was presented. This is a preliminary classification work on terms. It needs no training process and not even context information. Yet it obtains a relatively good result. The method itself is not domain dependent, thus it is applicable to different domains. Results show that in certain applications, a simple method may be more effective under similar circumstances. The algorithm can still be investigated over the use of more induction rules. Some context information, statistics of word/tag usage can also be explored.

### Acknowledgments

This project is partially supported by CERG grants (PolyU 5190/04E and PolyU 5225/05E) and B-Q941 (Acquisition of New Domain Specific Concepts and Ontology Update).

### References

- Yun Li, Qiangjun Wang. 2001. *Automatic Term Extraction in the Field of Information Technology*. In the proceedings of The Conference of 20th Anniversary for Chinese Information Processing Society of China.
- Yan He, Zhifang Sui, Huiming Duan, and Shiwen Yu. 2006. *Term Mining Combining Term Component Bank*. In *Computer Engineering and Applications*. Vol.42 No.33,4--7.
- Feng Zhang, Xiaozhong Fan, and Yun Xu. 2006. *Chinese Term Extraction Based on PAT Tree*. Journal of Beijing Institute of Technology. Vol. 15, No. 2.
- Van Rees, R. 2003. *Clarity in the Usage of the Terms Ontology, Taxonomy and Classification*. CIB73.

- Mikheev, A. 1997. *Automatic Rule Induction. for Unknown Word Guessing*. In Computational Linguistics Vol. 23(3), ACL.
- Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. In proceedings of HLT-NAACL.
- Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. *Morphological Features Help POS Tagging of Unknown Words across Language Varieties*. In proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.
- Samuelsson, Christer. 1993. *Morphological Tagging Based Entirely on Bayesian Inference*. In proceedings of NCCL 9.
- Brants, Thorsten. 2000. *TnT: A Statistical Part-of-Speech Tagger*. In proceedings of ANLP 6.
- G. Orphanos, and D. Christodoulakis. 1999. *POS Disambiguation and Unknown Word Guessing with Decision Trees*. In proceedings of EACL'99, 134--141.
- H Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In proceedings of International Conference on New Methods in Language Processing.
- Maosong Sun, Dayang Shen, and Changning Huang. 1997. *Cseg & Tag1.0: a practical word segmenter and POS tagger for Chinese texts*. In proceedings of the fifth conference on applied natural language processing.
- Ying Liu. 2002. *Analysing Chinese with Rule-based Method Combined with Statistic-based Method*. In Computer Engineering and Applications, Vol.7.
- Mikheev, A. 1996. *Unsupervised Learning of Word-Category Guessing Rules*. In proceedings of ACL-96.
- Nakagawa T, Kudoh T, and Matsumoto Y. 2001. *Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines*. In proceedings of NLP PPS 6, 325--331.
- Maosong Sun, Zhengping Zuo, and B K, TSOU. 2000. *Part-of-Speech Identification for Unknown Chinese Words Based on K-Nearest-Neighbors Strategy*. In Chinese Journal of Computers. Vol.23 No.2: 166--170.
- Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, Yirong Chen. 2007. *Chinese Terminology Extraction using Window-based Contextual Information*. In proceedings of CICLING.
- Huaping Zhang et al. 2003. *HHMM-based Chinese Lexical Analyzer ICTCLAS*. Second SIGHAN workshop affiliated with 41th ACL, 184--187. Sapporo Japan.
- Hui Wang. Last checked: 2007-08-04. *Statistical studies on Chinese vocabulary (汉语词汇统计研究)*. <http://www.huayuqiao.org/articles/wanghui/wanghui06.doc>. The date of publication is unknown from the online source.
- Ruifeng Xu, Qin Lu, Yin Li and Wanyin Li. 2005. *The Design and Construction of the PolyU Shallow Treebank*. International Journal of Computational Linguistics and Chinese Language Processing, V.10 N.3.
- Yunfang Wu, Baobao Chang and Weidong Zhan. 2003. *Building Chinese-English Bilingual Phrase Database*. Page 41-45, Vol. 4.





## **Technical Terminology in Asian Languages: Different Approaches to Adopting Engineering Terms**

### **Makiko Matsuda**

Nagaoka University of Technology  
1603-1 Kamitomioka, Nagaoka, Japan  
matsuda@vos.nagaokaut.ac.jp

### **Hiroki Goto**

University of Toyama  
Gofuku, Toyama, Japan  
hgoto@ctg.u-toyama.ac.jp

### **Robin Lee Nagano**

University of Miskolc  
Miskolc-Egyetemváros, Hungary  
nagano.robin@chello.hu

### **Tomoe Takahashi**

Nagaoka University of Technology  
1603-1 Kamitomioka, Nagaoka, Japan  
tomoe@kjs.nagaokaut.ac.jp

### **Yoshikazu Hayase**

Toyama National College of Maritime Technology  
1-2 Ebie-Neriya, Imizu, Japan  
hayase@toyama-cmt.ac.jp

### **Yoshiki Mikami**

Nagaoka University of Technology  
1603-1 Kamitomioka, Nagaoka, Japan  
mikami@kjs.nagaokaut.ac.jp

### **Abstract**

Terminology development in education, science and technology is a key to formulating a knowledge society. The authors are developing a multilingual engineering terminology dictionary consisting of more than ten thousand engineering terms each for ten Asian languages and English. The dictionary is primarily designed to support foreign students who are studying engineering subjects at Japanese higher educational institutions in the Japanese language. Analysis of the lexical terms could provide useful knowledge for language resource creators. There are two adoption approaches, “phonetic adoption” (transliteration of borrowed terms) and “semantic adoption” (where the meaning is expressed using native words). The proportion of the two options found in the terminology set of each country (or language) shows a substantial difference and seems to reflect language policies of the country and the influence of foreign languages on the host language. This paper presents preliminary results of our investigation on this question based on a comparative study of three languages: Japanese, Vietnamese and Thai.

### **1 Introduction**

#### **1.1 Terminology in Development Context**

For developing countries there is a strong necessity to establish an appropriate set of terminology to enable local people to learn science and technology in an efficient manner throughout their educational programs using their mother tongues as the instructional medium. Instruction in the mother tongue is beneficial for students to acquire a basic concept in a subject (UNESCO, 2003).

International communities are fully aware of this necessity. The World Summit on the Information Society (WSIS) addressed this issue, saying that terminology development in education, science and culture is a key to developing knowledge societies.

#### **1.2 Two Approaches to Adopting Terms**

Unlike the situation with basic vocabulary, scientific or technical terms are usually not “home-grown”, but are often imported from the outside world. And when a scientific or technical term is imported from one language (the source language) to another language (the host language), typically two approaches are found.

The first approach is to simply borrow a word by creating a phonetically equivalent word in the host language. For example, the equivalent of the

English word "science" in Malay is "sains", pronounced as /sains/, and "computer" in Japanese is written as "コンピュータ" in the Japanese katakana<sup>1</sup> syllabary and is pronounced /konpju:ta/. We call this **phonetic adoption** in this paper. It is quite similar to transliteration, but not exactly the same due to the difference in phonetic structure between the source and host languages.

The second approach is to create a new word by combining a semantically relevant word-root or word in the host language. For example, "science" in Thai is "วิทยาศาสตร์", pronounced /wít<sup>h</sup>ayasà:t/. This word is derived from the word "vidya" (knowledge) in Sanskrit, a source from which the Thai language has imported many words throughout its history. "Science" in Japanese is "科学", pronounced /kagaku/. This word was coined more than one hundred years ago as a combination of two Chinese characters, "科" and "学", meaning "a section, or a branch of something" and "learning", respectively. As shown in these two cases, it often happens that classical languages like Sanskrit or Chinese, rather than the host language itself, provide root-words in this creation process just as Latin and Greek roots are used in scientific and technical terms in English. We call this approach **semantic adoption** in this paper.

### 1.3 Pros and Cons of the Two Approaches

When we look back on the historical evolution of scientific terms dating from ancient civilization to modern times, both types of adoption are found. Both approaches to adoption have advantages and disadvantages in terms of ease of creation and ease of understanding, as well as other issues. The pros and cons of semantic adoption, summarized in Table 1, are in principle the opposites of those of phonetic adoption.

But the choice of adoption approach is not so simple. The phonetic adoption option is easy to implement, but would have the danger of disturbing language purity, which also has a high priority in many nations. When the semantic adoption option is chosen, language purity is maintained but the adoption cost is high and it takes time to pre-

pare a meaningful number of vocabulary items and to train teachers in their use.

Table 1. Pros and Cons of Phonetic and Semantic Approaches to Term Adoption

	Advantages	Disadvantages
Phonetic adoption	(1) easy to find connectivity to the original word (2) easy to create	(1) disturbs the purity of host language (2) not easy to guess what it means
Semantic adoption	(1) does not disturb the purity of host language (2) relatively easy to guess what it means, or at least to what it relates	(1) not easy to find connectivity to the original word (2) not easy to create

### 1.4 Third Approach: Use of foreign language

In addition to this dilemma, a completely different approach can be taken when the necessity is very urgent. This is to design and to implement the whole educational program by means of an established foreign language. We call this approach the **source language approach**. In the real world today, this third approach is widely adopted in many developing countries, to various extents. The choice of foreign language depends on subject domains and the economic and political condition of the country as well as on historical ties.

### 1.5 Objectives of the study

In summary, there are three basic options. However, little research has been carried out on how a language community can effectively formulate terminology focused on Asian languages. "Guidelines for Terminology Policy" (Infoterm 2005) talks about various issues relating to this subject, but no mention is found relating to the question of adoption strategy. The authors believe that an analysis of the existing terminology can provide practical and applicable lessons to terminology development practitioners and policy makers. This paper presents preliminary results of our investigation on this question, based on a comparative study of three countries: Japanese, Vietnamese and Thai.

## 2 Dictionary

In this paper, the main corpus is extracted from a multilingual engineering dictionary based on Babel (Hayase and Kawata 2002). Babel is a web tech-

<sup>1</sup> In Japanese, four scripts are used to write Japanese, e.g. *kanji* (i.e., Chinese characters), *hiragana*, *katakana* and the Latin alphabet. Katakana script is conventionally used to represent foreign phonetically translated words.

nology dictionary for foreign students in Japan<sup>2</sup>. It is composed of the 11 academic domains given in Table 2 and three languages, English, Japanese and Thai.

Table 2. The Number of Words in the Japanese Lexicon by Subject Domain

Subject Domain	Words (TYPE)	Words TOKEN
architecture	1,314	1,798
chemistry	717	958
civil engineering	766	1,263
telecommunications	650	1,410
computer science	941	1,538
control engineering	1,081	1,545
electronics	681	1,200
maritime science	961	1,403
mathematics	952	1,284
mechanical engineering	1,117	1,186
physics	587	945
others*	1,872	-
Total of all words	11,639	14,530

\*there is some overlap with other domains

Table 3. A Sample of the Lexicon

English	Japanese	Chinese	Vietnamese	Thai
discriminant	判別式	判別式	Biệt thức	วิธีแบ่งแยก ระหว่างกัน
factor theorem	因数定理	因数定理	Định lý thừa số	กฎบทตัวประกอบ
circular measure	弧度法	弧度法	Số đo cung tròn	การวัดเชิงวงกลม
radical root	累乘根	根	Căn	ราก
limit value	極限值	极限值	Giá trị giới hạn	ค่าจำกัด
divergence	発散	发散	Phát tán	การลู่ออก
natural logarithm	自然对数	自然对数	Đối số tự nhiên	ลอการิทึมธรรมชาติ
point of inflection	变曲点	拐点	Điểm uốn	จุดเปลี่ยนควมเว้า

To this primary source, we added some terms in mathematics and physics and made a multilingual database such as that shown in Table 3 by adding

<sup>2</sup> <http://www.toyama-cmt.ac.jp/%7Ehayase/Project/Babel/>

adoptions into 8 languages: Vietnamese, Chinese, Korean, Filipino, Malay, Sinhalese, Myanmar and Mongolian. The data are coded in Unicode. At present Vietnamese, Chinese, Korean and Sinhalese language have been translated and other languages will be finished soon. The dictionary will be made in April, 2008. For this paper, we have selected three languages for analysis – Vietnamese, Thai and Japanese – because these languages have relatively rich vocabularies.

### 3 Comparison of Adoption Approaches

The approach to terminology adoption is examined in this section from various angles.

#### 3.1 Comparison by Country

In order to investigate the portfolio of two adoption approaches, we surveyed the entire lexicon for all mechanical engineering subjects. In addition, we took a random sampling in our dictionary and chose approximately 25 words from two major subject fields: architecture and computer science. The portfolios of three languages are shown in Table 4. Compounds such as hybrids of native and loaned word-components were separated from semantic adoption.

Table 4. Comparison of Terms in 3 Domains

Subject Domain	# of terms	Adoption Approach	Origin of words	Rate		
				JP	VN	TH
mechanical engineering	1,186	Semantic	English	0.68	0.89	0.78
				0.32	0.09	0.21
				0.02	0.01	
computer science	23	phonetic	English	0.43	0.87	0.78
				0.57	0.13	0.21
				0.69	0.93	0.76
architecture	29	phonetic	English	0.31	0.03	0.24
				0.03	0.03	
				0.03		

Table 4 shows that the Japanese language has the highest percentage of phonetically translated words from English. Vietnamese has the lowest percentage of phonetic adopted words from English, but includes terms from French in mechanical engineering plus a few from Russian. Thai shows a

roughly similar balance between semantic adoption and English-origin terms in all three domains.

The high number of Thai and Vietnamese origin words in computer science is rather remarkable, as it indicates a consistent effort to adapt the Thai language to every emerging technology. Note that over half of the computer science terms for Japanese are of English origin. Vietnamese draws on different languages to varying degrees; it is notable that the languages of origin vary by subject field.

### 3.2 Comparison by Subject Domain: A Case of Japanese Terms

As we saw in Table 4, Japanese technical terms are often phonetically translated, but less so in architecture than in the other two fields. To gain a better idea of the distribution of phonetic adoption, we surveyed the entire Japanese lexicon for all subject domains. Table 5 gives the rate of katakana usage, that is, the rate of words written in the katakana script used for transcribing foreign words.

Table 5. Rate of katakana Usage

Subject Domain	# of terms	Katakana	Rate
architecture	1,798	147	0.081
chemistry	958	234	0.244
civil engineering	1,263	156	0.123
telecommunications	1,410	422	0.299
computer science	1,538	650	0.422
control engineering	1,545	497	0.321
electronics	1,200	250	0.208
maritime science	1,403	51	0.036
mathematics	1,284	50	0.038
mechanical engineering	1,186	385	0.324
physics	945	120	0.126

The domains of computer science, control engineering and mechanical engineering include over 30% of katakana words in their terms. In contrast, the percentage of katakana terms in architecture, mathematics and maritime science domains is less than 10%, indicating that these domains are highly semantically adopted in their terms, or have little need for imported vocabulary. We thus see that academic domain affects the adoption of phonetically translated terms.

### 3.3 Mathematical Terms at Different Grade Level

We have collected about 1,300 terms in mathematics for our multilingual engineering dictionary. Following the Japanese course of study<sup>3</sup> for grades 1-12, we chose 57 terms from our dictionary and compared those terms with their counterparts in Thai and Vietnamese. For university-level mathematical terms, following the syllabus of the 1<sup>st</sup> year of the Faculty of Mathematics in one Japanese University, we chose 11 terms<sup>4</sup>. Then we also surveyed those terms in the same way. Table 6 examines the terms used in various grade levels.

Table 6. Comparison of Mathematical Terms by Grade Level

Grade Level	# of terms	Origin of words	Rate		
			JP	VN	TH
University	11	Semantic (Sino)	0.91 (0.91)	0.90 (0.64)	0.64 (0.00)
		Phonetic	0.09	0.09	0.36
High School	17	Semantic (Sino)	1.00 (1.00)	1.00 (0.47)	0.88 (0.00)
		Phonetic	0.00	0.00	0.12
Junior high school	18	Semantic (Sino)	1.00 (1.00)	1.00 (0.61)	1.00 (0.00)
		Phonetic	0.00	0.00	0.00
Elementary school	22	Semantic (Sino)	1.00 (1.00)	1.00 (0.50)	1.00 (0.00)
		Phonetic	0.00	0.00	0.00

These data indicate that all languages use either original or semantically adopted from elementary level to junior-high school level. For Japanese almost all words are Sino-Japanese; in oral instruction native words may be used in Japan but those are not used as keywords. As for high school, some words phonetically translated from English are used in Thai, while at university level all languages use English-origin words. As for the rate of Sino-Vietnamese words in mathematics, it is considerably higher than other categories. This suggests that the likelihood that Sino-origin words can be shared

<sup>3</sup>The Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT) determines the national curriculum.

<sup>4</sup>We took terms from the 1<sup>st</sup> year syllabus of the Faculty of Mathematical Science in Doshisya University.

effectively in other countries using Chinese characters is generally high in the field of mathematics.

### 3.4 Web Presence of English Terms and their Adopted Equivalents

We also examined the web presence of pairs of terms, an English word and its equivalent in the host language. This is an attempt to find how widely the source language approach – direct use of foreign language – is used in technological topics. Four words from different subject domains were chosen, and a search was carried out for the term (as an exact phrase) using the Google search engine (Table 7).

Table 7. Web Presence of Technical Terms

Search through	Search term(s)	# of hits*
Japanese pages	EN ionization	121,000
	JP 電離	419,000
	EN fusion welding	387
	JP 融接	12,900
	EN automatic lathe	258
	JP 自動旋盤	56,600
	EN operand	32,500
	JP オペランド (katakana)	138,000
	JP 演算数 (kanji)	10,700
	EN ionization	89
Vietnamese pages	VN Điện ly	13,900
	EN fusion welding	6
	VN Hàn nung chảy	66
	EN automatic lathe	2
	VN Máy tiện tự động	48
	EN operand	111
	VN Số tính toán	810
	EN ionization	10,600
	TH การแตกประจุ	48
	EN fusion welding	692
Thai pages	TH การเชื่อมแบบหลอมละลาย	6
	EN automatic lathe	39
	TH เครื่องกลึงอัตโนมัติ	195
	EN operand	568
	TH ตัวถูกดำเนินการ	24

\*Accessed on Sept. 18, 2007

While data are limited to just four terms, general trends can be identified. For Japanese pages, while there were many hits for English, the hits for Japanese terms were at least three times higher. Interestingly, for *operand*, the one term with two equivalents, the katakana (i.e., phonetically adopted) term was used far more frequently than the semantically adopted kanji term. In the Vietnamese pages, while far fewer total hits were made, the Vietnamese terms were overwhelmingly more frequent than English terms. For Thai pages the opposite was found, with the exception of *automatic lathe*, a tool that may require marketing in the local language. These results suggest that the major language of technology for Thailand is English, for Vietnam it is the native language, and for Japan there is English use but Japanese dominates.

In order to find if there is a distinction between use in technological fields and in general use, we made a comparison using three non-technical terms (Table 8).

Table 8. Web Presence of Non-technical Terms

Search through	Search term(s)	# of hits*
Japanese pages	EN water	2,200,000
	JP 水	24,200,000
	EN novel	2,242,000
	JP 小説	9,870,000
	EN zoo	2,580,000
	JP 動物園	2,650,000
	EN water	295,000
	VN nước	4,310,000
	EN novel	68,400
	VN tiểu thuyết	2,130,000
Vietnamese pages	EN zoo	41,500
	VN vườn bách thú**	615,000
	VN sở thú***	1,750,000
	EN water	1,850,000
	TH น้ำ	2,610,000
	EN novel	269,000
	TH วรรณกรรม	1,830,000
	EN zoo	272,000
	TH สวนสัตว์	620,000

\*Accessed on Sept. 10, 2007

\*\*Northern dialect, \*\*\*Southern dialect

Again, we find that Japanese pages use the Japanese word far more frequently (although *zoo* is nearly equal – perhaps due to the exotic appeal of a foreign word, and one understood by almost every Japanese) but that English words have a sizable presence. For Vietnamese pages, the number of hits is substantially higher than that for the technical terms, but the same trend exists: the native language dominates. Finally, for Thai pages, the hits for Thai words far exceed those for English words in all three cases, although English words are often used.

A comparison of results for Tables 7 and 8 confirms that the role of languages differs between daily use and situations that require technical terminology, even in a country that has adopted the third approach of using a foreign language for science and technology.

## 4 Social and Academic Factors

### 4.1 High Rate of Semantic Adoption in Asian Countries

As can be imagined, it is not easy for Asian people to understand technical terms in other Asian languages, because semantic adoption is popular among Asian languages. The diversity of scripts in Asia is another reason of this difficulty. In European countries, term-sharing by phonetic adoption is easier because almost all languages were derived from the same origin and are sharing the same script. Therefore, in order to achieve mutual communication using technical terms in Asian countries, we need to make special efforts to go over the variation and harmonize terms each other.

But we need to discuss why those languages prefer semantic adoption. Various social, political, and historical factors influence language use. This applies also to the adoption of technical terminology. Here, we discuss influences of these factors on the approaches used to adopt technical terms, including the influence of technical domains and academic discipline.

### 4.2 Language Policies and Historical Background

#### 4.2.1 Vietnam

In Vietnam, adoption is strongly promoted under the strict language policy. The Political Resume of the Socialist Republic of Vietnam states that

“[e]very nationality has the right to use its own language and system of writing, to preserve its national identity, and to promote its fine customs, habits, traditions and culture.” Therefore they seldom use phonetically adopted words as such. One example of its limited use is “ôm” from the English “ohm.” Another example is proper nouns, when foreign words are used in newspapers. This political reason is the primary impetus for the semantic adoption of terms.

Secondly, as in other communist nations, English is not taught as an obligatory subject in Vietnam. This educational policy also brings about a low rate of phonetically adopted words from English, because many people cannot grasp the meaning.

Thirdly, we need to consider that the Vietnamese language has been historically influenced by several foreign languages, as we can induce from Table 4’s results. Vietnam was under the influence of China for approximately 1,000 years. A number of Chinese terms were absorbed into Vietnamese as Sino-Vietnamese words. During the French colonial period, the Vietnamese language added French words for the manufacture of specific products such as automobiles and bicycles. Lastly, at the end of the 20<sup>th</sup> century Russian technical terms were introduced by people who had studied in or technical advisers who had sent by the Soviet Union or in Russia

#### 4.2.2 Japan

As is the case with Vietnam, Chinese characters and words were imported from China so long ago that the characters and words are regarded as part of the native Japanese language. In the end of the 19<sup>th</sup> century, a number of Japanese-kango (words that used Chinese characters but were created by Japanese people) were invented as adoptions of western technical terms, and many of these terms were re-imported to China.

After World War II, Japan was occupied by the Allied Forces led by the United States and English became a compulsory subject from junior high school. English-origin words became easy to understand for many Japanese. This historical change caused a sharp rise in percentage of phonetic adoptions from English among Japanese technical terms (see Table 5) (Hashimoto 2007). However, English is not usually used as an instructional medium in

Japanese institutions of higher education. Teachers and students know many English loan terms but they use these loan terms not as English but as Japanese, and since they are pronounced according to Japanese phonetic rules, they may not even be intelligible to English speakers.

#### 4.2.2 Thai

Thai has different features from the other two countries in that it is less influenced by Chinese. But the Thai language has been influenced by another foreign language, Sanskrit, the classical Indian language. We did not examine the rate of Sanskrit origin words in the Thai language in this study, but it seems likely that a high rate of terms phonetically adopted from Sanskrit will be found in our engineering dictionary. Although Thai has native-language equivalents at the lexical level, English is daily used as an instructional medium in higher education. Students in Thailand start to learn English at 10 years old or earlier. Thus, even though they have native terms they prefer to use terms phonetically adopted from English or to use English.

### 4.3 Other Factors

#### 4.3.1 When the subject was introduced?

In the Meiji era (1868-1912) in Japan, experts translated technical terms from Western languages semantically, into Japanese-*kango*. Since World War II, however, it has been the mainstream to transliterate into katakana. Therefore a decisive factor in adoption approach is when the subject was introduced into Japan. Table 9 shows the foundation year for major academic societies and institutes in engineering fields, which roughly indicates the time of introduction of the subject.

As shown in Table 5, computer science is the field that uses technical terms in katakana most frequently, while the sampled vocabulary from telecommunications also includes nearly 30% of katakana words. These disciplines are relatively newly developed in Japan; the societies or institutes related to these fields were established in the latter half of the 20<sup>th</sup> century.

The oldest societies, the Mathematical Society of Japan and The Physical Society of Japan, were established in 1877; our study showed that only 3.8% of mathematical terms and 12.6% of terms in physics are phonetically adopted terms. Therefore,

disciplines with academic societies established relatively early may have developed a great deal of their terminology during the Meiji era, and may still have a preference for Japanese-*kango* over katakana. Newer disciplines, on the other hand, may have found it easier to use the imported terms in a relatively direct way.

Table 9. Academic Societies in Japan

Discipline	Major Academic Society in Japan	Est.
architecture	Architectural Institute of Japan	1886
civil engineering	Japan Society of Civil Engineers	1914
chemistry	Chemistry and Chemical Industry of Japan	1878
tele-communications, computer science	Information Processing Society of Japan	1960
	The Institute of Electronics, Information, and Communication Engineers	1987
	The Japan Society of Information and Communication Research	1983
control engineering	The Institute of Systems, Control and Information Engineers	1957
electronics	The Institute of Electrical Engineers of Japan	1888
maritime science	The Japan Society of Naval Architects and Ocean Engineers	1898
mathematics	Mathematical Society of Japan	1877
mechanical engineering	The Japan Society of Mechanical Engineers	1897
physics	The Physical Society of Japan	1877

#### 4.3.2. Is it indigenous?

Each country has specific characteristics of the region, such as regional climate, folk, culture, lifestyle and other indigenous factors. Technical terms in these fields are therefore strongly connected with regional characteristics. Before unfamiliar terms or new technologies were brought over to the host country, experts in those fields had already created and used their own technical terms in their native language, and also they were already aware of the concept of a thing indicated by an unfamiliar term in alien language. All they had to do was to make a one-to-one correspondence and correctly translate a foreign term into the term that was already in use in the mother tongue.

## 5 Conclusion

This preliminary study looks at the different approaches to adopting technical terms in three Asian languages. The three languages investigated prefer different methods: Vietnamese tends to adopt words into its language semantically, through adoption, while Japanese adopts them phonetically, through transliteration. Thai, to a large extent, has chosen a third approach, that of using a source language (English).

In Japanese and in Vietnamese, many terms subject to semantic adoption were translated into Chinese characters. This suggests that Sino-words may be an effective way of communicating concepts among certain language users in certain disciplines, such as mathematics.

The effect of grade level was also investigated, and it was found that native language terms were used in mathematics almost exclusively until high school or university level.

Subject domain was found to have an effect on the adoption approach and was often influenced by whether the domain had a long tradition, and when established a discipline was (as shown by the foundation of academic societies). A domain effect was seen in Vietnamese, where Russian or French-origin terms appeared in different domains.

While quite limited in scope, this study has revealed clear trends that deserve further investigation.

## 6 Future Research

We plan to investigate more fully the rate of phonetically translated words and approach to terminology adoption. We expect that Chinese and Korean will give us further evidence of widespread use of Sino-words and Mongolian will provide data for phonetic adoption from Russian. Malay and Filipino are likely to be highly influenced by English, while we will find the influence of Sanskrit in Myanmar and Sinhalese.

## Acknowledgements

The study was made possible by the sponsorship of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) through the Asian Language Resource Network project. The authors would like to express sincere gratitude to all lexicon contributors for their help.

## References

- Christian Galinski. 1995. Terminology and Standardization under a machine adoption perspective. *MT Summit V Proceedings*.
- Douglas Skuce, Ingrid Meyer. 1990. *Concept Analysis and Terminology: A Knowledge-Based Approach to Documentation, International Conference on Computational Linguistics archive Proceedings of the 13th conference on Computational linguistics*, Volume 1.
- Giovanni Battista Varile, Antonio Zampolli. 1997. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.
- Hashimoto Waka. 2007. A diachronic transition of phonetically translated words in Japanese: based on the column of newspaper. *Language*, 36-6, Taishukan Publishers, Japan. (橋本和佳 「外来語の通時的推移 - 新聞社説を素材として - 」 『言語』 36-6, 大修館書店)
- Hayase Yoshikazu, Shigeo Kawata. 2002. BABEL: A multilingual dictionary and hypertext formation system for engineers. *Transactions of JSCES*, Paper No.20020023.(早勢欣和, 川田重夫「BABEL: エンジニア支援のための多言語辞書とハイパーテキスト生成」)
- Infoterm. 2005. *UNESCO Guidelines for Terminology Policies*. Formulating and implementing terminology policy in language communities, Paris.
- Kageura Kyo. 2006. The status of borrowed items in Japanese terminology. *Studies in Japanese Language* Vol2, No.4 (影浦峯 「日本語専門語彙の構成における外来語語基の位置づけ」 『日本語の研究』 2006)
- UNESCO. 2003. Education in a multilingual world. Paris: UNESCO, (ED-2003/WS/2)  
<http://unesdoc.unesco.org/images/0012/001297/129728e.pdf>



# Selection of XML tag set for Myanmar National Corpus

**Wunna Ko Ko**

AWZAR Co.

Mayangone Township, Yangon,  
Myanmar

wunnakoko@gmail.com

**Thin Zar Phyo**

Myanmar Unicode and NLP Research  
Center

Myanmar Info-Tech, Hlaing Campus,  
Yangon, Myanmar

myanmar.nlp5@gmail.com

## Abstract

In this paper, the authors mainly describe about the selections of XML tag set for Myanmar National Corpus (MNC). MNC will be a sentence level annotated corpus. The validity of XML tag set has been tested by manually tagging the sample data.

Keywords: Corpus, XML, Myanmar, Myanmar Languages

## 1 Introduction

Myanmar (formerly known as Burma) is one of the South-East Asian countries. There are 135 ethnic groups living in Myanmar. These ethnic groups speak more than one language and use different scripts to present their respective languages. There are a total of 109 languages spoken by the people living in Myanmar [Ethnologue, 2005].

There are seven major languages, according to the speaking population in Myanmar. They are Kachin, Kayin/Karen, Chin, Mon, Burmese, Rakhine and Shan [Ko Ko & Mikami, 2005]. Among them, Burmese is the official language and spoken by about 69% of the population as their mother tongue [Ministry of Immigration and Population, 1995].

Corpus is a large and structured set of texts. They are used to do statistical analysis, checking occurrences or validating linguistic rules on a specific universe.<sup>1</sup>

In Myanmar, there are a plenty of text for most of the languages, especially Burmese and major languages, since stone inscription.

Myanmar Language Commission and a number of scholars had been collected a number of corpora for their specific uses [Htay et al., 2006]. But there is no national corpus collection, both in digital and non-digital format, until now.

Since there are a number of languages used in Myanmar, the national level corpus to be built will include all languages and scripts used in Myanmar. It has been named as Myanmar National Corpus or MNC, in short form.

During the discussion for the selection of format for the corpus, XML (eXtensible Markup Language), a subset of SGML (Standard Generalized Markup Language), format has been chosen since XML format can be a long usable and possible to keep the original format of the text [Burnard, 1996]. The range of software available for XML is increasing day by day. Certainly more and more NLP related tools and resources are produced in it. This in turn makes the necessity of selection of XML tag set to start building of MNC.

MNC will include not only written text but also spoken texts. The part of written text will include regional and national newspapers and periodicals, journals and interests, academic books, fictions, memoranda, essays, etc. The part of spoken text will include scripted formal and informal conversations, movies, etc.

During the selection of XML tag sets, the sample for all the data which will be included in building of MNC, has been learnt.

## 2 Myanmar National Corpus

Myanmar is a country of using 109 different languages and a number of different scripts [Ethnologue, 2005]. In order to do language processing for these languages and scripts, it becomes a necessity to build a corpus with

<sup>1</sup> [http://en.wikipedia.org/wiki/Text\\_corpus](http://en.wikipedia.org/wiki/Text_corpus)

languages and scripts used in Myanmar; at least with major languages and scripts, which will include almost all areas of documents.

Among the different scripts used in Myanmar, the popular scripts include Burmese script (a Brahmi based script), Latin scripts. Building of MNC will be helpful for development of Natural Language Processing (NLP) tools (such as grammar rules, spelling checking, etc) and also for linguistic research on these languages and scripts. Moreover, since Burmese script is written without necessarily pausing between words with spaces, the corpus to be built is hoped to be useful for developing tools for automatic word segmentation.

## 2.1 XML based corpus

XML is universal format for structured documents and data, and can provide highly standardized representation frameworks for NLP (Jin-Dong KIM et al. 2001); especially, the ones with annotated corpus based approaches, by providing them with the knowledge representation frameworks for morphological, syntactic, semantics and/or pragmatics information structure. Important features are:

- XML is extensible and it does not consist of a fixed set of tags.
- XML documents must be well-formed according to a defined syntax.
- XML document can be formally validated against a schema of some kind.
- XML is more interested in the meaning of data than its presentation.

The XML documents must have exactly one top-level element or root element. All other elements must be nested within it. Elements must be properly nested [Young, 2001]. That is, if an element starts within another element, it must also end within that same element.

Each element must have both a start-tag and an end-tag. The element type name in a start-tag must exactly match the name in the corresponding end-tag and element name are case sensitive.

Moreover, the advantages of XML for NLP includes ontology extraction into XML based structured languages using XML Schema. The

great benefit about XML is that the document itself describes the structure of data.<sup>2</sup>

Three characteristics of XML distinguish from other markup languages:<sup>3</sup>

- its emphasis on descriptive rather than procedural markup;
- its notion of documents as instances of a *document type* and
- its independence of any hardware or software system.

Since MNC is to be built in XML based format, the selection process for tag set of XML become an important process. The XML tagged corpus data should also keep the original format of the data.

In order to select XML tag set for MNC, the sample data for the corpus has to be collected. The format of the sample corpus data has been studied for the selection of the XML tag set in appropriate with the data format.

## 2.2 Structure of a data file at MNC

The structure of a data file at MNC will include two main parts: information of the corpus file and the corpus data.

The first part, the header part of a corpus file, describes the information of a corpus file. The information of the corpus file includes the header which will provide sensible use of the corpus information in machine readable form. In this part, the information such as language usage and the description of the corpus file will be included.

The second part, the document part, of a corpus file will include the source description of the corpus data and the corpus data, the written or spoken part of the text, itself. The information of the corpus data such as bibliographic information, authorship, and publisher information will be included in this section. Moreover, the corpus data itself will also be included in this section.

The hierarchically structure of a corpus file at MNC will be as shown in figure 1.

<sup>2</sup> <http://www.tei-c.org/P5/Guidelines/index.html>

<sup>3</sup> <http://www.w3.org/TR/xml/>

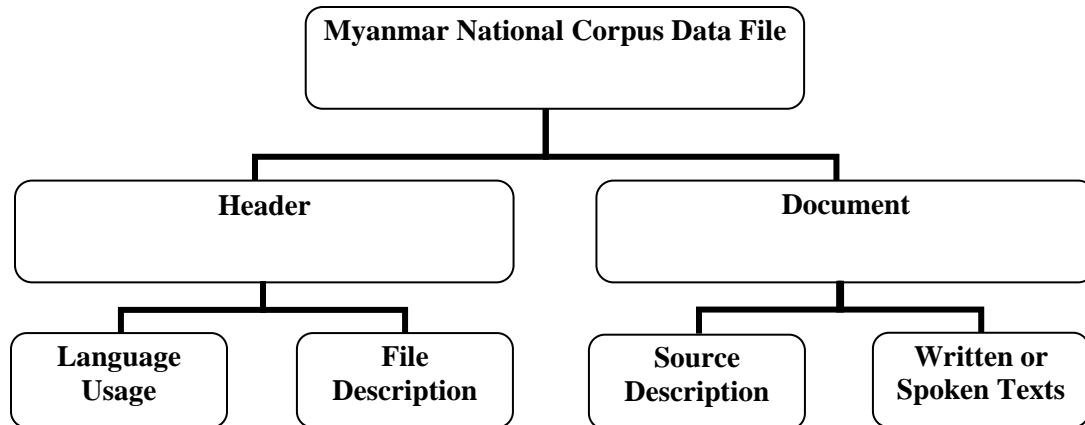


Figure 1. Hierarchically structure of a data file at MNC

### 3 Selection of necessary XML tag set

After studying original formats and features of texts, to be used in corpus, and the structure of the corpus file has been determined, the selection procedure for XML tag set has been started.

British National Corpus (BNC)<sup>4</sup>, American National Corpus (ANC)<sup>5</sup> had been referenced for selection of XML tag set.

The selection of XML tag set is based on the nature of the structure of a data file. The main tag for the data file will be named as <mnc> which is the abbreviation of Myanmar National Corpus.

A data file contains two main parts, the header part and the document part.

```

-<mnc>
+<teiHeader></teiHeader>
+<myaDoc></myaDoc>
</mnc>
  
```

Figure 2. Root and element tags of MNC

#### 3.1 Header Part

The XML tag for the header part of the corpus data file is named as <teiHeader>. Text Encoding Initiative (TEI) published guidelines

<sup>4</sup> Lou Burnard. 2000. Reference Guide for the British National Corpus (World Edition). Oxford University Computing Services, Oxford.

<sup>5</sup> Nancy Ide and Keith Suderman. 2003. The American National Corpus, first Release. Vassar College, Poughkeepsie, USA

for the text encoding and Interchange<sup>6</sup>. TEI encoding scheme consists of a number of rules with which the document has to adhere in order to be accepted as a TEI document.

This header part contains language usage of the data file <langUsage> and the file description <fileDesc> which includes machine readable information of the data file.

```

-<mnc>
-<teiHeader>
+<langUsage></langUsage>
+<fileDesc></fileDesc>
</teiHeader>
+<myaDoc></myaDoc>
</mnc>
  
```

Figure 3. Element and Child tags of MNC

The language usage part contains such information as language name <langName>, script information <script>, International Organization for Standardization (ISO) code number <ISO>, encoding information <encodingDesc> and version of encoding <version>.

```

-<mnc>
-<teiHeader>
<langUsage>
<langName> </langName>
<script> </script>
<ISO></ISO>
<encodingDesc> </encodingDesc>
  
```

<sup>6</sup> TEI Consortium. 2001, 2002 and 2004 Text Encoding Initiative. In The XML Version of the TEI Guidelines.

```

    <version> </version>
  </langUsage>
  +<fileDesc></fileDesc>
</teiHeader>
+<myaDoc></myaDoc>
</mnc>

```

**Figure 4. 2<sup>nd</sup> level Child tags in language Usage part of MNC**

The file description part contains such information as title information of the corpus file <titleStmt>, edition information <editionStmt> and publication information about the corpus file <publicationStmt>. The detail information will be tagged using more specific lower level child tags under the previously described tags.

```

-<mnc>
  -<teiHeader>
    +<langUsage></langUsage>
  -<fileDesc>
    +<titleStmt></titleStmt>
    +<editionStmt></editionStmt>
    +<publicationStmt></publicationStmt>
  </fileDesc>
</teiHeader>
+<myaDoc></myaDoc>
</mnc>

```

**Figure 5. 2<sup>nd</sup> level Child tags in file description part of MNC**

### 3.2 Document Part

The XML tag for the document part of the corpus data file is named as <myaDoc> which is the short form of Myanmar Document. It contains two sub parts: the source description of the data <sourceDesc> and the original data itself which in turn can be divided into two types; written text <wtext> and the spoken text <stext>.

```

<mnc>
  +<teiHeader></teiHeader>
  -<myaDoc>
    +<sourceDesc></sourceDesc>
    +<wtext></wtext>
  </myaDoc>
</mnc>

```

**Figure 6. Element and Child tags of MNC**

The first part, the source description part of the data <sourceDesc>, will contain the

bibliographic information, such as title, name of author, publisher, etc., of the original data.

```

<mnc>
+<teiHeader></teiHeader>
-<myaDoc>
  -<sourceDesc>
    -<bibl>
      <title></title>
      <author></author>
      <editor/></editor>
    -<imprint>
      <publisher></publisher>
      <pubPlace></pubPlace>
      <date></date>
    </imprint>
    </bibl>
  </sourceDesc>
  +<wtext></wtext>
</myaDoc>
</mnc>

```

**Figure 7. 2<sup>nd</sup> level Child tags for source description part of MNC**

The second part, the original data part <wtext> or <stext> will contain the whole original data. The original format information such as heading <head type="MAIN">, sub-heading <head type="SUB">, paragraph number <paragraph n="1">, sentence number <s n="1"> will be saved in this part.

```

<mnc>
  +<teiHeader></teiHeader>
  -<myaDoc>
    +<sourceDesc></sourceDesc>
    -<wtext>
      -<head>
        <s></s>
        +<paragraph></paragraph>
        +<head></head>
      </head>
    </wtext>
  </myaDoc>
</mnc>

```

**Figure 8. 2<sup>nd</sup> level Child tags for original data part of MNC**

Since MNC is going to be annotated in sentence level, each sentence will be annotated and numbered.

```

<mnc>
+<teiHeader></teiHeader>
-<myaDoc>
+<sourceDesc></sourceDesc>
-<wtext>
  -<head>
    <s></s>
  -<paragraph>
    -<s></s>
    </paragraph>
  </head>
+<head></head>
</wtext>
</myaDoc>
</mnc>

```

**Figure 9. Down to the sentence level Child tags of MNC**

### 3.3 Sample MNC data file

The Myanmar National Corpus is a major resource for linguistic research, as well as computational linguistics research, lexicography, corpus linguistic research and a resource for the development of Myanmar Language teaching material because we expect the corpus to be continually expanded in the future.

A sample MNC data is use the Universal Declaration of Human Rights (UDHR) texts in Burmese and Karen, which is one of the major languages in Myanmar, has been used to sample tagging with the selected XML tag set.

The following figure is show for the sample MNC.

```

<? xml version="1.0"?>
<mnc>
-<teiHeader>
  -<langUsage>
    <langName> Myanmar </langName>
    <script>Burmese</script>
    <ISO> 10646</ISO>
    <encodingDesc> utf-8</encodingDesc>
    <version>Unicode 5.0</version>
  </langUsage>
-<fileDesc>
  -<titleStmt>
    <title>Myanmar National Corpus</title>
  -<respStmt>
    <resp>Corpus built by</resp>
    <name>Myanmar NLP Team</name>
  </respStmt>
  </titleStmt>
-<editionStmt>
  <edition> First TEI-conformant version </edition>
  <extent/>
</editionStmt>
-<publicationStmt>
  <address>Myanmar Info-Tech, Yangon, Myanmar</address>
  <availability status="restricted">
    Availability limited to Myanmar NLP Team
  </availability>
-<creation>
  <date>07/06/2007</date>
</creation>
  <distributor>Myanmar NLP Team </distributor>
  <idno type="mnc">MNC101</idno>

```

```

        </publicationStmt>
    </fileDesc>
</teiHeader>

-<myaDoc xml:id="TEXTS">
    -<sourceDesc>
        -<bibl>
            <title>
                အပြည်ပြည်ဆိုင်ရာလူ့အခွင့်အရေးကြေညာစာတမ်း
                (meaning: Universal Declaration of Human Rights)
            </title>
            <author/>
            <editor/>
            -<imprint vol="64" n="46">
                <publisher></publisher>
                <pubPlace></pubPlace>
                <date></date>
            </imprint>
        </bibl>
    </sourceDesc>

    -<wtext type="OTHERPUB">
        -<head type="MAIN">
            <s n="1">
                အပြည်ပြည်ဆိုင်ရာလူ့အခွင့်အရေးကြေညာစာတမ်း
                (meaning: Universal Declaration of Human Rights)
            </s>
            +<paragraph n="1"></paragraph>
            -<head type="SUB">
                <s n="1"> စကားချိုး (meaning: Preamble) </s>
                -<paragraph n="1">
                    -<s n="1">
                        လူ့ခပ်သိမ်း၏ မျိုးရိုးဂုဏ်သိက္ခာနှင့် တကွလူတိုင်းအညီအမျှခံစားခွင့်ရှိသည့်အခွင့်အရေးများကို
                        အသိအမှတ်ပြုခြင်းသည်လူ့ခပ်သိမ်း၏ လွတ်လပ်မှု၊ တရားမျှတမှု၊ ငြိမ်းချမ်းမှုတို့၏ အခြေခံအုတ်မြစ်
                        ဖြစ်သောကြောင့်လည်းကောင်း၊ .....
                        (meaning: Whereas recognition of the inherent dignity and of the equal and
                        inalienable rights of all members of the human family is the foundation of
                        freedom, justice and peace in the world,.....)
                    </s>
                </paragraph>
                +<paragraph n="2"></paragraph>
            </head>
            -<head type="SUB">
                <s n="2"> အပိုဒ် ၁ (meaning: paragraph 1) </s>
                -<paragraph n="1">
                    <s n="1">
                        လူတိုင်းသည်တူညီလွတ်လပ်သောဂုဏ်သိက္ခာဖြင့်လည်းကောင်း၊ တူညီလွတ်လပ်သောအခွင့်အရေး
                        များဖြင့်လည်းကောင်း၊ မွေးဖွားလာသူများဖြစ်သည်။
                        (meaning: All human beings are born free and equal in dignity and rights.)
                    </s>
                </paragraph>
            </head>
        </wtext>
    </myaDoc>

```

```

        </s>
        <s n="2">
        ထိုသူတို့၌ပိုင်းခြားဝေဖန်တတ်သောဉာဏ်နှင့်ကျင့်ဝတ်သိတတ်သောစိတ်တို့ရှိကြ၍ထိုသူတို့သည်
        အချင်းချင်းမေတ္တာထား၍ဆက်ဆံကျင့်သုံးသင့်၏။
        (meaning: They are endowed with reason and conscience and should act towards
        one another in a spirit of brotherhood.)
        </s>
    </paragraph>
    -<head type="SUB">
        <s n="3"> အပိုဒ် ၂ (meaning: paragraph 2) </s>
        +<paragraph n="1"></paragraph>
        +<paragraph n="2"></paragraph>
    </head>
    -<head type="SUB">
        <s n="4"> အပိုဒ် ၃ (meaning: paragraph 2) </s>
        -<paragraph n="1">
            <s n="1">
                လူတိုင်း၌အသက်ရှင်ရန်လွတ်လပ်မှုနှင့်လုံခြုံစိတ်ချခွင့်ရှိသည်။
                (meaning: Everyone has the right to life, liberty and security of person.)
            </s>
        </paragraph>
    </head>
    +<head type="SUB"></head>
    +<head type="SUB"></head>
</head>
</wtext>
</myaDoc>
</mnc>

```

Figure 10. Sample MNC Corpus file (Burmese UDHR text in MNC XML format)

#### 4 Conclusion and Future work

In this paper, the authors have clearly described about the selection of XML tag set for building of MNC. Since the word level segmentation for Burmese script is not yet available, the corpus data will be annotated only up to the sentence level in order to be in the same format for all Myanmar languages and scripts.

In order to check whether the selected the XML tag set will be enough and useful for tagging the corpus data, the sample corpus data has been collected by manually tagging the data which includes newspapers and periodicals, Universal Declaration of Human Rights (UDHR), novels and essays.

Since the manual tagging to the sample corpus data proves that the selected XML tag set is enough to cover a variety of data sources, the

next step is to develop an algorithm for automatic tagging the data.

#### Acknowledgement

This study was performed with the support of the Government of the Union of Myanmar through Myanmar Natural Language Implementation Committee. Thanks and gratitude towards the members of Myanmar Language Commission for providing necessary information to write this paper.

#### References

Ethnologue. 2005 *Languages of the World*, 15<sup>th</sup> Edition, Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com/>. Edited by Raymond G. Gordon, Jr.

- Hla Hla Htay, G. Bharadwaja Kumar and Kavi N. Murthy. 2006. *Constructing English-Myanmar Parallel Corpora*. The Fourth International Conference on Computer Application 2006 (ICCA 2006) Conference Program.
- Jin-Dong KIM, Tomoko OHTA, Yuka TATEISI, Hideki MIMA and Jun'ichi TSUJII. 2001. *XML-based Linguistic Annotation of Corpus*. In the Proceedings of the first NLP and XML Workshop held at NLPRS 2001. pp. 47--53.
- Lou Burnard. 1996. *Using SGML for Linguistic Analysis: the case of the BNC*. ACM Vol 1 Issue 2 (Spring 1999) MIT Press ISSN: 1099-6621. pp. 31-51.
- Michael J. Young. 2001. *Step by Step XML*. Prentice Hall of India Private Limited Press. ISBN-81-203-1804-B
- Ministry of Immigration and Population. 1995. *Myanmar Population Changes and Fertility Survey 1991*. Immigration and Population Department
- Wunna Ko Ko, Yoshiki Mikami. 2005 *Languages of Myanmar in Cyberspace*, In Proceedings of TALN & RECITAL 2005 (NLP for Under-Resourced Languages Workshop), Dourdan, FRANCE, 2005 June, pp. 269-278.



# Myanmar Word Segmentation using Syllable level Longest Matching

**Hla Hla Htay, Kavi Narayana Murthy**

Department of Computer and Information Sciences

University of Hyderabad, India

hla\_hla\_htay@yahoo.co.uk, knmuh@yahoo.com

## Abstract

*In Myanmar language, sentences are clearly delimited by a unique sentence boundary marker but are written without necessarily pausing between words with spaces. It is therefore non-trivial to segment sentences into words. Word tokenizing plays a vital role in most Natural Language Processing applications. We observe that word boundaries generally align with syllable boundaries. Working directly with characters does not help. It is therefore useful to syllabify texts first. Syllabification is also a non-trivial task in Myanmar. We have collected 4550 syllables from available sources. We have evaluated our syllable inventory on 2,728 sentences spread over 258 pages and observed a coverage of 99.96%. In the second part, we build word lists from available sources such as dictionaries, through the application of morphological rules, and by generating syllable n-grams as possible words and manually checking. We have thus built list of 800,000 words including inflected forms. We have tested our algorithm on a 5000 sentence test data set containing a total of (35049 words) and manually checked for evaluating the performance. The program recognized 34943 words of which 34633 words were correct, thus giving us a Recall of 98.81%, a Precision of 99.11% and a F-Measure is 98.95%.*

**Key Words:-** Myanmar, Syllable, Words, Segmentation, Syllabification, Dictionary

## 1 Introduction

Myanmar (*Burmese*) is a member of the Burmese-Lolo group of the Sino-Tibetan language spoken by about 21 Million people in Myanmar (Burma). It is a tonal language, that is to say, the meaning of a syllable or word changes with the tone. It has been classified by linguists as a mono-syllabic or isolating language with agglutinative features. According to history, Myanmar script has originated from *Brahmi* script which flourished in India from about 500 B.C. to over 300 A.D (MLC, 2002). The script is syllabic in nature, and written from left to right.

Myanmar script is composed of 33 consonants, 11 basic vowels, 11 consonant combination symbols and extension vowels, vowel symbols, devow- elizing consonants, diacritic marks, specified symbols and punctuation marks(MLC, 2002),(Thu and Urano, 2006). Myanmar script represents sequences of syllables where each syllable is constructed from consonants, consonant combination symbols (i.e. Medials), vowel symbols related to relevant consonants and diacritic marks indicating tone level.

Myanmar has mainly 9 parts of speech: noun, pronoun, verb, adjective, adverb, particle, conjunction, post-positional marker and interjection (MLC, 2005), (Judson, 1842).

In Myanmar script, sentences are clearly delimited by a sentence boundary marker but words are not always delimited by spaces. Although there is a general tendency to insert spaces between phrases, inserting spaces is more of a convenience rather than

a rule. Spaces may sometimes be inserted between words and even between a root word and the associated post-position. In fact in the past spaces were rarely used. Segmenting sentences into words is therefore a challenging task.

Word boundaries generally align with syllable boundaries and syllabification is therefore a useful strategy. In this paper we describe our attempts on syllabification and segmenting Myanmar sentences into words. After a brief discussion of the corpus collection and pre-processing phases, we describe our approaches to syllabification and tokenization into words.

Computational and quantitative studies in Myanmar are relatively new. Lexical resources available are scanty. Development of electronic dictionaries and other lexical resources will facilitate Natural Language Processing tasks such as Spell Checking, Machine Translation, Automatic Text summarization, Information Extraction, Automatic Text Categorization, Information Retrieval and so on (Murthy, 2006).

Over the last few years, we have developed *monolingual text corpora* totalling to about 2,141,496 sentences and *English-Myanmar parallel corpora* amounting to about 80,000 sentences and sentence fragments, aligned at sentence and word levels. We have also collected word lists from these corpora and also from available dictionaries. Currently our *word list* includes about 800,000 words including inflected forms.

## 2 Myanmar Words

Myanmar words are sequences of syllables. The syllable structure of Burmese is C(G)V((V)C), which is to say the onset consists of a consonant optionally followed by a glide, and the rhyme consists of a monophthong alone, a monophthong with a consonant, or a diphthong with a consonant<sup>1</sup>. Some representative words are:

- CV [mei] girl
- CVC [me ' ] crave
- CGV [mjei] earth
- CGVC [mje ' ] eye

<sup>1</sup>[http://en.wikipedia.org/wiki/Burmese\\_language](http://en.wikipedia.org/wiki/Burmese_language)

- CVVC [maun] (term of address for young men)
- CGVVC [mjau] ditch

Words in the Myanmar language can be divided into simple words, compound words and complex words (Tint, 2004),(MLC, 2005),(Judson, 1842). Some examples of compound words and loan words are given below.

- Compound Words

- head [u:] ဦး + pack [htou ' ] ထုပ် = hat [ou ' htou ' ] ဦးထုပ်
- language [sa] စာ + look,see [kji.] ကြည့် + [tai ' ] building တိုက် = library [sa kji. dai ' ] စာကြည့်တိုက်
- sell [yaun:] ရောင်း + buy [we] ဝယ် = trading ရောင်းဝယ် [*yaun : we*]

- Loan Words

- ကွန်ပျူတာ [kun pju ta] computer
- ဆင်ကော်မတီ [hsa ' ko mā ti] sub-committee
- ချယ်ရီ [che ri] cherry

## 3 Corpus Collection and Preprocessing

Development of lexical resources is a very tedious and time consuming task and purely manual approaches are too slow. We have downloaded Myanmar texts from various web sites including news sites including official newspapers, on-line magazines, trial e-books (over 300 full books) as well as free and trial texts from on-line book stores including a variety of genres, types and styles - modern and ancient, prose and poetry, and example sentences from dictionaries. As of now, our corpus includes 2,141,496 sentences.

The downloaded corpora need to be cleaned up to remove hypertext markup and we need to extract text if in pdf format. We have developed the necessary scripts in Perl for this. Also, different sites use different font formats and character encoding standards are not yet widely followed. We have mapped these various formats into the standard *WinInnwa* font format. We have stored the cleaned up texts in ASCII format and these pre-processed corpora are seen to be reasonably clean.

## 4 Collecting Word Lists

Electronic dictionaries can be updated much more easily than published printed dictionaries, which need more time, cost and man power to bring out a fresh edition. Word lists and dictionaries in electronic form are of great value in computational linguistics and NLP. Here we describe our efforts in developing a large word list for Myanmar.

### 4.1 Independent Words

As we keep analyzing texts, we can identify some words that can appear independently without combining with other words or suffixes. We build a list of such valid words and we keep adding new valid words as we progress through our segmentation process, gradually developing larger and larger lists of valid words. We have also collected from sources such as Myanmar Orthography(MLC, 2003), CD versions of English-English-Myanmar (Student's Dictionary)(stu, 2000) and English-Myanmar Dictionary (EMd, ) and Myanmar-French Dictionary (damma sami, 2004). Currently our word list includes 800,000 words.

### 4.2 Stop Word Removal

Stop words include prepositions/post-positions, conjunctions, particles, inflections etc. which appear as suffixes added to other words. They form closed classes and hence can be listed. Preliminary studies therefore suggested that Myanmar words can be recognized by eliminating these stop words. Hopple (Hopple, 2003) also notices that particles ending phrases can be removed to recognize words in a sentence. We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of English-English-Myanmar (Student's Dictionary)(stu, 2000), English-Myanmar Dictionary (EMd, ) and The Khit Thit English-Myanmar dictionary (Saya, 2000). We have also looked at stop word lists in English (www.syger.com, ) and mapped them to equivalent stop words in Myanmar. See Table 1. As of now, our stop words list contains 1216 entries. Stop words can be prefixes of other stop words leading to ambiguities. However, usually the longest matching stop word is the right choice.

Identifying and removing stop words does not

Nominative personal pronouns	
I	ကျွန်တော် [kjun do], ကျွန်မ [kja ma.], ငါ [nga], ကျုပ် [kjou '], ကျနော် [kja no], ကျုပ် [kjanou '], ကျမ [kja ma.]
Possessive pronouns and adjectives	
my	ကျွန်ုပ်၏ [kjou ' i.], ကျွန်တော်၏ [kjun do i.], ကျွန်မ၏ [kja ma. i.], ကျနော်၏ [kja nou ' i.], ကျမ၏ [kja ma. i.], ငါ့ [nga i.], ကျုပ်ရဲ့ [kjou ' i.], ကျွန်ုပ်ရဲ့ [kjou ' je.], ကျွန်တော်ရဲ့ [kjun do je.], ကျွန်မရဲ့ [kja ma. je.], ကျနော်ရဲ့ [kja nou ' je.], ကျမရဲ့ [kja ma. je.], ငါ့ရဲ့ [nga je.], ကျုပ်ရဲ့ [kjou ' je.], ကျွန်တော် [kjun do.], ကျနော် [kja no.]
Indefinite pronouns and adjectives	
some	အချို့ [a chou.], အချို့သော [a chou. tho.], တချို့ [ta chou.], တချို့သော [a chou. tho:], တချို့ချို့ [ta chou.ta chou.], တချို့တလေ [ta chou.ta lei]

Table 1: Stop-words of English Vs Myanmar

always necessarily lead to correct segmentation of sentences into words. Both under and over segmentation are possible. When stop-words are too short, over segmentation can occur. Under segmentation can occur when no stop-words occur between words. Examples of segmentation can be seen in Table 2. We have observed that over segmentation is more frequent than under segmentation.

ဝိုင်းဝန်းမျိုးကျားစံရသဖြင့်သူအနေကံသည်		
ဝိုင်းဝန်းမျိုးကျားစံရ	အနေကံ	
[waing: win: chi: kyu: khan ya]	[a nay khak]	
received compliments	abashed	
V <sub>pp</sub>	V <sub>past</sub>	
ကျောင်းအုပ်ဆရာကြီးသည်အကြမ်းဖက်မှုကိုခက်ဆပ်သည်		
ကျောင်းအုပ်ဆရာကြီး	အကြမ်းဖက်မှု	ခက်ဆပ်
[kyaung: aop hsa ya kyi:]	[a kyan: phak mhu]	[sak sop]
The headmaster	violence	abhors
N <sub>subj</sub>	N <sub>obj</sub>	V <sub>present</sub>

Table 2: Removing stop-words for segmentation

### 4.3 Syllable N-grams

Myanmar language uses a syllabic writing system unlike English and many other western languages which use an alphabetic writing system. Interestingly, almost every syllable has a meaning in Myanmar language. This can also be seen from the work of Hopple (Hopple, 2003).

Myanmar Natural Language Processing Group has listed 1894 syllables that can appear in Myanmar texts (Htut, 2001). We have observed that there are more syllables in use, especially in foreign words including Pali and Sanskrit words which are widely used in Myanmar. We have collected other pos-

sible syllables from the Myanmar-English dictionary(MLC, 2002). Texts collected from the Internet show lack of standard typing sequences. There are several possible typing sequences and corresponding internal representations for a given syllable. We include all of these possible variants in our list. Now we have over **4550** syllables.

Bigram bisyllables	Trigram 3-syllables	4-gram 4-syllables
လန်အိမ် lantern [hpan ein]	ပုန့်ခန့် with a big sound [boun: gə ne:]	နှစ်နှစ်ကာကာ whole-heartedly [hni ' hni ' ka ga]
ပန်သာ: glassware [hpan tha:]	ရွှေ့ခန့် effortlessly [swei. gə ne:]	ထူးထူးကဲကဲ outstanding [htu: htu: ke: ke:]
ကန်စောင်း: bank of lake [kan saun:]	ထောင်းခန့် fuming with rage [htaun: gə ne:]	များများစားစား many,much [mja: mja: sa: za:]

Table 3: Examples of Collected N-grams

No. of syllables	No of words	Example
1	4550	ကောင်း Good (Adj) [kaun:]
2	59964	လိပ်ပြာ Butterfly, Soul (N) [lei ' pja]
3	170762	ပြတင်းပေါက် Window (N) [b ə din: bau ' ]
4	274775	ပြည်ထွင်းထုတ်ကုန် Domestic Product (N) [pji dwin: htou ' koun]
5	199682	လျှပ်စစ်ထမင်းအိုး [hlja ' si ' ht ə min: ou:] Rice Cooker(N)
6	99762	သူမပြုဆရာမ Nurse(female) (N) [thu na bju. hs ə ja ma.]
7	41499	ရင်းနှီးသူ:ကြပေတော့သည် become friend (V) [jin: hni: thwa: kya. pei to. thi]
8	14149	ပြည်ထောင်စုမြန်မာနိုင်ငံတော် Union of Myanmar (N) [pji daun zu. mj ə ma nain gan to ]
9	4986	သယံဇာတအရင်းအမြစ်များ Natural Resources (N) [than jan za ta. ə jin: ə mji ' ]
10	1876	မြေကိုခါခါလှုပ်စေခြင်း be agitated or shaken(V) [ chei ma kain mi. le ' ma kain mi. hpji ' thi]

Table 4: Syllable Structure of Words

We have developed scripts in Perl to syllabify words using our list of syllables as a base and then generate n-gram statistics using Text::Ngrams which is developed by Vlado Keselj (Keselj, 2006). This program is quite fast and it took only a few minutes on a desktop PC in order to process 3.5M bytes

of Myanmar texts. We have used “-type=word” option treating syllables as words. We had to modify this program a bit since Myanmar uses zero (as “(0) wa ” letter) and the other special characters ( “;”, “<”, “>”, “:”, “&”, “[”, “]” etc.) which were being ignored in the original Text::Ngrams software. We collect all possible words which is composed of n-grams of syllables up to 5-grams. Table 1 shows some words which are collected through n-gram analysis. Almost all monograms are meaningful words. Many bi-grams are also valid words and as we move towards longer n-grams, we generally get less and less number of valid words. See Table 3. Further, frequency of occurrence of these n-grams is a useful clue. See Table 4.

By analyzing the morphological structure of words we will be able to analyze inflected and derived word forms. A set of morphemes and morphological forms have been collected from (MLC, 2005) and (Judson, 1842) . See Table 5. For example, the four-syllable word in Table 3 is an adverb “ထူးထူးကဲကဲ” [htu: htu: ke: ke:] outstanding derived from the verb “ထူးကဲ”. See Table 3.

Statistical construction of machine readable dictionaries has many advantages. New words which appear from time to time such as Internet, names of medicines, can also be detected. Compounds words also can be seen. Common names such as names of persons, cities, committees etc. can be also mined. Once sufficient data is available, statistical analysis can be carried and techniques such as mutual information and maximum entropy can be used to hypothesize possible words.

#### 4.4 Words from Dictionaries

Collecting words using the above three mentioned methods has still not covered all the valid words in our corpus. We have got only 150,000 words. Words collected from n-grams needs exhaustive human effort to pick the valid words. We have therefore collected words from two on-line dictionaries - the English-English-Myanmar (Student’s Dictionary) (stu, 2000), English-Myanmar Dictionary (EMd, ) and from two e-books - French-Myanmar(damma sami, 2004), and Myanmar Orthography (MLC, 2003). Fortunately, these texts can be transformed into winninnwa font. We have

A basic unit 1 syllable	B (Verb)= A + သည်	C (Noun)= အ+A	D (Negative)= မ+A+ဘူး	E (Noun)= A+မှု
ကောင်း [kaun:] good (Adj)	ကောင်းသည် [kaun: thi] is good	အကောင်း [a kaun:] good	မကောင်းဘူး [ma. kaun: bu:] Not good	ကောင်းမှု [kaun: mhu.] good deeds
ဆိုး [hso:] bad (Adj)	ဆိုးသည် [hso: thi] is bad	အဆိုး [a hso:] bad	မဆိုးဘူး [ma. hso: bu:] Not bad	ဆိုးမှု [hso: mhu.] Bad Deeds
ရောင်း [jaun:] sell(Verb)	ရောင်းသည် [jaun: thi] sell	အရောင်း [a jaun:] sale	မရောင်းဘူး [ma. jaun: bu:] not sell	ရောင်းမှု [jaun: mhu.] sale
ရေး [jei:] write(Verb)	ရေးသည် [jei: thi] write	အရေး [a jei:] writing	မရေးဘူး [ma. jei: bu:] do not write	ရေးမှု [jei: mhu.] writing
ပြော [pjo:] talk,speak(Verb)	ပြောသည် [pjo: thi] talk,speak	အပြော [a pjo:] talk,speech	မပြောဘူး [ma. pjo: bu:] not talk,speak	ပြောမှု [pjo: mhu.] talking

Table 5: Example patterns of Myanmar Morphological Analysis

written Perl scripts to convert to the standard font. Myanmar Spelling Bible lists only lemma (root words). We have suffixed some frequently used morphological forms to these root words.

There are lots of valid words which are not described in published dictionaries. The entries of words in the Myanmar-English dictionary which is produced by the Department of the Myanmar Language Commission are mainly words of the common Myanmar vocabulary. Most of the compound words have been omitted in the dictionary (MLC, 2002). This can be seen in the preface and guide to the dictionary of the Myanmar-English dictionary produced by Department of the Myanmar Language Commission, Ministry of Education. 4-syllables words like “ ထူးထူးဆန်းဆန်း: ”[htu: htu: zan: zan:] (strange), “ ထူးထူးကဲကဲ ” [htu: htu: ke: ke:](outstanding) and “ ထူးထူးခြားခြား: ” [htu: htu: gja: gja:](different)(see Table 3) are not listed in dictionary although we usually use those words in every day life.

With all this, we have been able to collect a total of about 800,000 words. As we have collected words from various sources and techniques, we believe we have fairly good data for further work.

On screen	ကြီး	ကို
In ascii	MuD:	udk
	BuD:	ukd

Table 6: Syllables with different typing sequences

## 5 Syllabification and Word Segmentation

Since dictionaries and other lexical resources are not yet widely available in electronic form for Myanmar language, we have collected 4550 possible syllables including those used in Pali and foreign words such as ဓမ္မတက္ကတိလံ ), considering different typing sequences and corresponding internal representations, and from the 800,000 strong Myanmar word-list we have built. With the help of these stored syllables and word lists, we have carried out syllabification and word segmentation as described below. Many researchers have used longest string matching (Angell et al., 1983),(Ari et al., 2001) and we follow the same approach.

The first step in building a word hypothesizer is syllabification of the input text by looking up syllable lists. In the second step, we exploit lists of words (viewed as n-grams at syllable level) for word segmentation from left to right.

### 5.1 Syllabification

As an initial attempt we use longest string matching alone for Myanmar text syllabification. Examples are shown in Table 7.

**Pseudo code** Here we go from left-to-right in a greedy manner:

```

sub syllabification{
  Load the set of syllables from syllable-file
  Load the sentences to be processed from sentence-file
  Store all syllables of length j in N_j where j = 10..1
  for-each sentence do
    length ← length of the sentence

```

```

pos ← 0
while (length > 0) do
  for j = 10..1 do
    for-each syllable in Nj do
      if string-match sentence(pos, pos + j) with syllable
        Syllable found. Mark syllable
        pos ← pos + j
        length ← length - j
      End if
    End for
  End for
End while
Print syllabified string
End for
}

```

```

for j = 10..1 do
  for-each word in Nj do
    if string-match sentence(pos, pos + j) with word
      word found. Mark word
      pos ← pos + j
      length ← length - j
    End if
  End for
End for
End while
Print tokenized string
End for

```

We have evaluated our syllables list on a collection of 11 short novels entitled “*Orchestra*” ခ်-ဝဲဝဲဝဲဝဲ:[than zoun ti: wain:], written by “**Nikoye**” (Ye, 1997) which includes 2,728 sentences spread over 259 pages including a total of 70,384 syllables. These texts were syllabified using the longest matching algorithm over our syllable list and we observed that only 0.04% of the actual syllables were not detected. The Table 6 shows that different typing sequences of syllables were also detected. Here are some examples of failure: ခ်ဝဲဝဲ:[rkdCf;]and ခ်ဝဲဝဲ:[rkdvf;] which are seldom used in text. The typing sequence is also wrong. Failures are generally traced to

- differing combinations of writing sequences
- loan words borrowed from foreign languages
- rarely used syllables not listed in our list

## 5.2 Word Segmentation

We have carried out tokenization with longest syllable word matching using our 800,000 strong stored word list. This word list has been built from available sources such as dictionaries, through the application of morphological rules, and by generating syllables n-grams and manually checking. An example sentence and its segmentation is given in Table 8.

```

Load the set of words from word-file
for-each word do
  i ← syllabification(word);
  Store all words of syllable length i in Ni where i = 10..1
End for

```

```

Load the sentences to be processed from sentence-file
for-each sentence do
  length ← syllabification(sentence);
  #length of the sentence in terms of syllables
  pos ← 0
  while (length > 0) do

```

## 6 Evaluation and Observations

We have segmented 5000 sentences including a total of (35049 words) with our programs and manually checked for evaluating the performance. These sentences are from part of the English-Myanmar parallel corpus being developed by us (Htay et al., 2006). The program recognized 34943 words of which 34633 words were correct, thus giving us a Recall of 98.81% and a Precision of 99.11%. The F-Measure is 98.95%. The algorithm suffers in accuracy in two ways:

**Out-of-vocabulary Words:** Segmentation error can occur when the words are not listed in dictionary. No lexicon contains every possible word of a language. There always exist out-of-vocabulary words such as new derived words, new compounds words, morphological variations of existing words and technical words (Park, 2002). In order to check the effect of out-of-vocabulary words, we took a new set of 1000 sentences (7343 words). We have checked manually and noticed 329 new words, that is about 4% of the words are not found in our list, giving us a coverage of about 96%.

### Limitations of left-to-right processing:

Segmentation errors can also occur due to the limitations of the left-to-right processing. See the example 1 in Table 9. The algorithm suffers most in recognizing the sentences which have the word *He* ခ် [thu] followed by a *negative verb* starting with the particle ဝဲ [ma.]. The program wrongly segments *she* as *he*. Our text collection obtains from various sources and the word “she” is used as ခ်ဝဲ [thu ma.] in modern novels and Internet text. Therefore, our

ကော်မီသောက်ရင်းအနံ့တို့နှင့်အလပသလပပြောနေခဲ့သည်																
aumfzDaomuf&fif;tefwDESihftvyovyajymaecJhonf																
ကော်	မီ	သောက်	ရင်း	အနံ့	တို့	နှင့်	အ	လ	ပ	သ	လ	ပ	ပြော	နေ	ခဲ့	သည်
aumf	zD	aomuf	&fif;	tef	wD	ESihf	t	v	y	o	v	y	ajym	ae	cJh	onf
[ko]	[hpi]	[thau ']	[jin:]	[an]	[ti]	[hnin.]	[a]	[la]	[pa.]	[tha.]	[la]	[pa.]	[pjo]	[nei]	[khe.]	[thi]

Table 7: Example syllabification

ကျောင်းအုပ်ဆရာကြီးသည်အကြမ်းဖက်မှုကိုစက်ဆုပ်သည်				
ကျောင်းအုပ်ဆရာကြီး	သည်	အကြမ်းဖက်မှု	ကို	စက်ဆုပ်သည်
[kyaung: aop hsa ya ky:]	[thi]	[a kyan: phak mhu]	[ko]	[sak sop thi]
The headmaster		violence		abhors
N <sub>subj</sub>	Particle	N <sub>obj</sub>	Particle	V <sub>present</sub>

Table 8: A sentence being segmented into words

word list contains she သူမ. This problem can be solved by standardization. Myanmar Language Commission (MLC, 1993) has advised that the words “she” and “he” should be written only as သူ and the word သူမ representing a feminine pronoun should not be used. For example 2 in Table 9, the text အားပေးသည် can be segmented into two ways. 1) အားပေးသည် [a: pei: thi] which means “encourage” and 2) အား: [particle for indicating dative case] and ပေးသည် give [pei: thi]. Because of greedy search from left to right, our algorithm will always segment as အားပေးသည် no matter what the context is.

In order to solve these problems, we are plan to use machine learning techniques which 1) can also detect real words dynamically (Park, 2002) while we are segmenting the words and 2) correct the greedy cut from left to right using frequencies of the words from the training samples.

Although our work presented here is for Myanmar, we believe that the basic ideas can be applied to any script which is primarily syllabic in nature.

## 7 Conclusions

Since words are not uniformly delimited by spaces in Myanmar script, segmenting sentences into words is an important task for Myanmar NLP. In this paper we have described the need and possible techniques for segmentation in Myanmar script. In particular, we have used a combination of stored lists, suffix removal, morphological analysis and syllable level n-grams to hypothesize valid words with about 99% accuracy. Necessary scripts have been written in Perl. Over the last few years, we have col-

lected monolingual text corpora totalling to about 2,141,496 sentences and English-Myanmar parallel corpora amounting to about 80,000 sentences and sentence fragments, aligned at sentence and word levels. We have also built a list of 1216 stop words, 4550 syllables and 800,000 words from a variety of sources including our own corpora. We have used fairly simple and intuitive methods not requiring deep linguistic insights or sophisticated statistical inference. With this initial work, we now plan to apply a variety of machine learning techniques. We hope this work will help to accelerate work in Myanmar language and larger lexical resources will be developed soon.

## References

Richard C. Angell, George W. Freurd, and Peter Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19(4):255–261.

Pirkola Ari, Heikki Keskustalo, Erkkka Leppnen, Antti-Pekka Knsl, and Kalervo Jrvelin. 2001. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2):235–237, january.

U damma sami. 2004. *Myanmar-French Dictionary*.

English-myanmar dictionary. Ministry of Education, Union of Myanmar,CD version.

Paulette Hopple. 2003. *The structure of nominalization in Burmese,Ph.D thesis*. May.

Hla Hla Htay, G. Bharadwaja Kumar, and Kavi Narayana Murthy. 2006. Building english-myanmar parallel corpora. In *Fourth International Conference on Computer Applications*, pages 231–238, Yangon, Myanmar, Feb.

Example 1: ဓားပြမှုတွင်သူမပါဝင်ခဲ့ဘူး					
ဓားပြမှု	တွင်	သူမ	ပါဝင်ခဲ့ဘူး		
[damja. hmu.]	[twin]	[thu ma.]	[pa wun khe. bu:]		
robbery	in	she	did not involve		
N	Particle	N <sub>subj</sub>	V <sub>pastneg</sub>		
Example 2: မိမိမလိုချင်သောတာဝန်ကိုသူတစ်ပါးအားပေးသည်။					
မိမိ	မလိုချင်သော	တာဝန်	ကို	သူတစ်ပါး	အားပေးသည်
[mi. mi.]	[ma. lou chin tho:]	[ta wun]	[gou]	[thu daba:]	[a: pei: thi]
I,myself	don't want	duty,responsibility		others	encourage
N <sub>subj</sub>	V <sub>neg</sub>	N <sub>obj1</sub>	Particle	N <sub>obj2</sub>	V

Table 9: Analysis of Over-Segmentation

Zaw Htut. 2001. All possible myanmar syllables, September.

Adoniram Judson. 1842. *Grammatical Notices of the Buremse Langauge*. Maulmain: American Baptist Mission Press.

Vlado Keselj. 2006. Text ::ngrams. <http://search.cpan.org/~vlado/Text-Ngrams-1.8/>, November.

MLC. 1993. *Myanmar Words Commonly Misspelled and Misused*. Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar.

MLC. 2002. *Myanmar-English Dictionary*. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

MLC. 2003. *Myanmar Orthography*. Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar, June.

MLC. 2005. *Myanmar Grammer*. Department of the Myanmar Language Commission, Ministry of Education,Union of Myanmar, June.

Kavi Narayana Murthy. 2006. *Natural Language Processing - an Information Access Perspective*. Ess Ess Publications, New Delhi, India.

Youngja Park. 2002. Identification of probable real words : an entropy-based approach. In *ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

U Soe Saya. 2000. *The Khit Thit English-English-Myanmar Dictionary with Pronunciation*. Yangon, Myanmar, Apr.

2000. Student's english-english/myanmar dictionary. Ministry of Commerce and Myanmar Inforithm Ltd, Union of Myanmar, CD version, Version 1, April.

Ye Kyaw Thu and Yoshiyori Urano. 2006. Text entry for myanmar language sms: Proposal of 3 possible input methods, simulation and analysis. In *Fourth International Conference on Computer Applications*, Yangon, Myanmar, Feb.

U Tun Tint. 2004. Features of myanmar language. May. [www.syger.com](http://www.syger.com). <http://www.syger.com/jsc/docs/stopwords/english.htm>.

Ni Ko Ye. 1997. *Orchestra*. The two cats, June.



# The Link Structure of Language Communities and its Implication for Language-specific Crawling

**Rizza Camus Caminero**

Language Observatory  
Nagaoka University of Technology  
Nagaoka, Niigata, Japan  
rhyze.caminero@gmail.com

**Yoshiki Mikami**

Language Observatory  
Nagaoka University of Technology  
Nagaoka, Niigata, Japan  
mikami@kjs.nagaokaut.ac.jp

## Abstract

Since its inception, the World Wide Web (WWW) has exponentially grown to shelter billions of monolingual and multilingual web pages that can be navigated through hyperlinks. Its structural properties provide useful information in presenting the socio-linguistic properties of the web. In this study, about 26 million web pages under the South East Asian country code top-level-domains (ccTLDs) are analyzed, several language communities are identified, and the graph structure of these communities are analyzed. The distance between language communities are calculated by a distance metrics based on the number of outgoing links between web pages. Intermediary languages are identified by graph analysis. By creating a language subgraph, the size and diameter of its strongly-connected components are derived, as these values are useful parameters for language-specific crawling. Performing a link structure analysis of the web pages can be a useful tool for socio-linguistic and technical research purposes.

## 1 Introduction

The World Wide Web contains interlinked documents, called web pages, navigable by hyperlinks. Since its creation, it has grown to contain billions of web pages and several billion hyperlinks. These web pages are created by millions of people from all parts of the world. Each web page contains a

large amount of information that can be shared and disseminated to people with Internet access. Authors and creators of a web page come from different backgrounds, different cultures, and different languages. Thus, a web page is a resource of multilingual content, and a fertile source for socio-linguistic analysis.

### 1.1 Web Crawling

While search-engines are important means of accessing the web for most users, automated systems of retrieving information from the web have been developed. These systems are called web crawlers, where a software agent is given a list of pages to visit. As the crawler visits these pages, it follows their outgoing links and adds these to the list of pages to visit. Each page is visited recursively according to some sets of policies (e.g. the type of pages to retrieve, direction, the maximum depth from the URL (uniform resource locator), etc.). The result of this crawl is a vast amount of data, most of which may be irrelevant to certain individuals. Thus, a focused-crawling approach was implemented by some systems to limit the search to only a subset of the web.

Focused crawlers rely on classifiers to work effectively. Language-specific crawlers, for example, need a very good language identification module that properly identifies the language of the web pages. General crawlers can be extended to include focused-crawling capabilities by incorporating the classifiers. Another important requirement to efficiently crawl the desired domain is the list of initial web pages, called seed URLs. Each of these URLs will be enqueued into a list. The agent visits each URL in the list. Since the crawler recursively visits

the outgoing links of each URL, it is possible that the seed URL is an outgoing link of another seed URL, or an outgoing link of an outgoing link of another seed URL. Listing these URLs as seed URLs will just waste the crawler's time in visiting them, when they have already been visited. If several URLs can be reached from just one URL, the seed URL list size will decrease and the crawler will be more efficient. However, the maximum distance between these web pages must also be considered, since the crawler can have a policy to stop the crawling after reaching a certain depth.

## 1.2 The Web as Graph

A graph consists of a set of nodes, denoted by  $V$  and a set of edges, denoted by  $E$ . Each edge is a pair of nodes  $(u, v)$  representing a connection between  $u$  and  $v$ . A path between two nodes is a sequence of edges that is passed through from one node to reach the other node.

In a directed graph, each edge is an ordered pair of nodes. A path from  $u$  to  $v$  does not imply a path from  $v$  to  $u$ . The distance between any two nodes is the number of edges in a shortest path connecting them. The diameter of the graph is the maximum distance between any two nodes.

In an undirected graph, each edge is an unordered pair of nodes. There is an edge between  $u$  and  $v$  if there is a link between  $u$  and  $v$ , regardless of which node is the source of the link.

A strongly-connected component (*SCC*) of a directed graph is a set of nodes such that for any pair of nodes  $u$  and  $v$  in the set, there is a path from  $u$  to  $v$ . The strongly-connected components of a graph consist of disjoint sets of nodes. A *subgraph* is a graph whose nodes and edges are subsets of another graph.

With the interlinked nature of the web, it can be represented as a graph, with the web pages as the nodes and the edges as the hyperlinks between the pages.

## 1.3 Languages

Languages are expressions of individuals and groups of individuals, essential to all form of communication. It is a fundamental medium of expressing one's self whether in spoken or written form. Ethnologue (2005) lists 6,912 known living languages in the world. Only a small portion of these languages can be found in the web today.

A language community in the web is the group of web pages written in a language. Major language communities discovered in each country indicates the dominant language of the country's web space. How one language community is related to another language community can be shown by analyzing the hyperlinks between them. Thus, a language graph can be created with the language communities as nodes, and the links between the language communities as edges.

## 2 Previous Studies

One of the earliest web survey in Asia (Ciolek, 1998) presented statistical data of the Asian web space by using the Altavista WWW search engine in gathering its data. In 2001, he wrote a paper presenting the trends in the volume of hyperlinks connecting websites in 10 East Asian countries.

Several studies have also been done regarding the representation of the web as a graph. Kumar et al. (2000) showed that a graph can be induced by the hyperlinks between pages. Measures on the connected component sizes and diameter were presented to show the high-level structure of the web. Broder et al. (2000) did experiments on the web on a larger scale and showed the web's macroscopic structure consisting of the SCC, IN, OUT, and TENDRILS. Balakrishnan and Deo (2006) observed that the number of edges grow superlinearly with the number of nodes, showing the degree distributions and diameter. Petricek et al. (2006) used web graph structural metrics to measure properties such as the distance between two random pages and interconnectedness of e-government websites. Bharat et al. (2001) studied the macro-structure of the web, showing the linkage between web sites by creating the "hostgraph", with the nodes representing the hosts and the edges as the count of hyperlinks between pages on the corresponding hosts.

Chakrabarti et al. (1999) proposed a new approach to topic-specific web resource discovery by creating a focused crawler that selectively retrieves web pages relevant to a pre-defined set of topics. Stamatakis et al. (2003) created CROSSMARC, a focused web crawler that collects domain-specific web pages. Deligenti et al. (2000) presented a focused crawling algorithm that builds a model for the context within which relevant pages to a topic occur on the web. Pingali et al. (2006) created an Indian search engine with a language identification

module that returns a language only if the number of words in a web page are above a given threshold value. The web pages were transliterated first into UTF-8 encoding. Tamura et al. (2007) presented a simulation study of language specific crawling and proposed a method for selectively collecting web pages written in a specific language by doing a linguistic graph analysis of real web data, and then transforming them into variation of link selection strategies.

Despite several studies on web graph and language-specific crawling, no study has been done showing the “language graph”. Herring et al. (2007) showed a study on language networks of a selected web community, LiveJournal, but not on the web as a whole.

### 3 Scope and Objectives of the Study

This research was conducted on the 10 ccTLDs of the South East Asian countries. This paper, however, will only show the results for the Indonesian domain (.id).

This research aims to show the socio-linguistic properties of the language communities in each country at the macroscopic level. The web page distribution for each language community in a given ccTLD and its most frequently linked to languages are shown. The distance is also computed and the language graph is illustrated.

This research also aims to show the graph properties of some Filipino language communities and its implication for crawling. Graph properties like the SCC size and the diameter will be presented to show these characteristics in a subset of the web.

Finally, this research demonstrates the usefulness of graph analysis approaches.

## 4 Methodology

This study was conducted by performing a series of steps from the collection of data to the presentation of the results through images.

### 4.1 UbiCrawler

UbiCrawler (Boldi et al., 2004) was used to download the web pages under the Asian ccTLDs. These pages were downloaded primarily for the purpose of assessing the usage level of each language in cyberspace, one of the objectives of

the Language Observatory Project (LOP)<sup>1</sup>. The crawl was started on July 5, 2006, running for 14 days. 107,168,733 web pages were collected from 43 ccTLDs in Asia. Each page contains several information such as the character set and outgoing links. For this study, the URL of a web page and the URL of its outgoing links were used. Although there are many web pages of Asia that can be found in generic domains, they were not included in this survey to limit the volume of the crawl data.

### 4.2 Language Identification

A language identification module (LIM) developed for LOP based on the n-gram method (Suzuki et al., 2002) was used to identify on what language a page is written in. This method first creates byte sequences for each training text of a language. It then checks the byte sequences of the web pages that match the byte sequences of the training texts. The language having the highest matching rate is considered as the language of the web page. The language identification module used the parallel corpus of the Universal Declaration of Human Rights (UDHR), translated into several languages. After crawling, LIM was executed to identify the languages of each downloaded web page. The identification result was stored in a LIM result file that contains the URL, the language, and matching rate, among others. In this study, the issues regarding the accuracy of LIM will not be discussed.

### 4.3 Web Page Analysis

For this study, the web pages of the 10 South East Asian ccTLDs were selected for analysis. There were 26,196,823 web pages downloaded under these ccTLDs. The web pages for each country were grouped by languages. The list of languages was narrowed down to 20 based on the number of pages, arranged from highest to lowest.

The link structure can be analyzed by traversing the outgoing links of each web page. For each web page, its outgoing links are retrieved. For each outgoing link, the LIM result file is checked for its language. The number of outgoing links in each language is counted. If the URL of the outgoing link is not on the file, it wasn't downloaded. Therefore, the language of the outgoing link is unidentified. This is usually the case of outgoing links un-

<sup>1</sup> <http://www.language-observatory.org>

der the generic TLDs (e.g., .com, .org, .gov, etc.) and non-Asian ccTLDs.

#### 4.4 Language Graph

There is a link between two languages if there is at least one outgoing link from a web page in one language to the other language. The language graph is created through contraction procedure, where all edges linking the same language page are contracted.

#### 4.5 Language Adjacency Matrix

Based on the number of web pages in a language and the number of outgoing links from one language to another, the language adjacency table  $N$  for each country is created. The row and column headers are the same – the top 20 languages based on the number of web pages. The value  $N_{ij}$  is the number of outgoing links from language  $i$  to language  $j$ .

The language adjacency matrix  $P$  contains the ratio of the number of outgoing links and the total number of outgoing links as can be found in the language adjacency table. Each cell value,  $P_{ij}$  is the probability that a web page in a language  $i$  has an outgoing link to language  $j$ .

$$P_{ij} = N_{ij} / \sum_k N_{ik}$$

A link from language  $i$  to language  $j$  is not necessarily accompanied by a link from language  $j$  to  $i$ . Even if there is a link, the number of outgoing links is not equal. To show the relationship between two languages based on the link structure, the language distance is computed.

#### 4.6 Distance between Languages

The distance between two languages measures their level of connectedness. It is the relationship between the number of outgoing links from language  $i$  to language  $j$  and vice versa. The distance is computed as the ratio of the number of outgoing links between two languages and the total number of outgoing links of the two languages.

The distance between language  $i$  and language  $j$ ,  $D_{ij}$  is,  $D_{ij} = \sqrt{1/(R_{ij} + \alpha)} - \beta$  where  $R_{ij}$  is the language link ratio is defined as,

$$R_{ij} = (N_{ij} + N_{ji}) / \left( \sum_k N_{ik} + \sum_k N_{jk} \right) \quad \text{for } (i \neq j)$$

$$R_{ij} = 1 \quad \text{for } (i = j)$$

where  $\alpha$  is an adjusting parameter introduced to avoid division-by-zero, which may happen when  $R_{ij}=0$ , i.e. no links between two languages. We set  $\alpha=0.0001$ . Thus, the maximum distance between languages becomes 99.  $\beta$  is another adjusting parameter to make  $D_{ij}$  as a distance metrics, and we set  $\beta=1$ . Assumption behind this definition is based on commonly-observed rules in our world. It is widely observed that interaction between two objects is proportional to the inverse square of distance between two objects. The number of web links between two language communities is considered as a kind of interaction. Languages with no links between them have a distance of 99, and the distance of a language to itself is 0.

Based on this distance metrics, the macroscopic language graph is created. A distance limit of 15 is used to clearly show which languages are closely-related by their link structure.

#### 4.7 Intermediary Languages

Considering the direction of the outgoing links, the possibility that language  $i$  will link to language  $j$  may be lesser than the possibility that language  $i$  will link to language  $j$  by passing through an intermediary language  $k$ , such that  $P_{ik}P_{kj} > P_{ij}$ . The intermediary language is identified, as this would mean that there are better ways to reach another language from one language.

#### 4.8 Graph Analysis using JGraphT

JGraphT is a free Java graph library that provides graph objects and algorithms. The library provides classes that calculates and returns the strongly-connected components subgraphs. To compute the distance between nodes, several graph searching algorithms are available, one of which is the Dijkstra algorithm that computes for the shortest path. A utility to export the graph into a format readable by most graph visualization tools is also available. A graph file, written in the DOT language (a plain text graph description language) was created, containing the nodes and edges of language pages. From this, the strongly-connected components and its properties (i.e. size, diameter) were determined.

#### 4.9 GraphVis: Visualization Tool

GraphViz is open-source graph visualization software that takes descriptions of graphs in a simple text language and makes diagrams in

several formats, including images. The neat layout, which makes “spring model” layouts, was used to visualize the distance between languages. However, the calculated distance cannot be drawn exactly, and this visualization is only two-dimensional. So, the images are distorted and do not illustrate the exact distance, only an approximation.

## 5 Results

This section shows some results on the Indonesian domain.

### 5.1 Link Structure

Indonesia is a country with one of the biggest language diversity in the world. According to Ethnologue<sup>2</sup>, 742 languages are spoken in the country. But, the LIM results show that only five of these languages are listed in the top 10 languages in the country, i.e., Javanese, Indonesian, Malay, Sundanese, and Madurese.

No.	Language	# of Pages	# of Outgoing Links
1	Javanese (jav)	797,300 (28.01%)	33,411,032 (27.20%)
2	English (eng)	743,457 (26.11%)	16,645,014 (13.55%)
3	Indonesian (ind)	516,528 (18.14%)	20,783,793 (16.92%)
4	Thai (tha)	218,453 (7.67%)	8,952,101 (7.29%)
5	Malay (mly)	197,535 (3.47%)	4,990,402 (4.06%)
6	Sundanese (sun)	98,835 (3.47%)	5,349,194 (4.35%)
7	Luxemburg <sup>3</sup> (ltz)	43,376 (1.52%)	2,307,602 (1.88%)
8	Occitan <sup>4</sup> (inc)	27,663 (0.97%)	351,318 (0.29%)
9	Madurese (mad)	22,121 (0.78%)	777,903 (0.63%)
10	Tatar (tat)	20,709 (0.73%)	3,334,651 (2.71%)
	Others	160,917 (5.65%)	25,930,885 (21.11%)
	Total	2,846,894	122,833,898

Table 1. LIM result for Indonesian Domain

<sup>2</sup> Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International.

<sup>3</sup> Luxembourgish

<sup>4</sup> Occitan Languidocien

Javanese is the most popular language in the web space of Indonesia, and constitutes 28% of the total number of web pages. It is followed by English and Indonesian. Although Indonesian is an official language of the country, it is ranked third.

English, a major business language of Indonesia, has the second largest number of web pages in the domain. But, Indonesian occupies the second rank in the number of outgoing links.

No.	Language	Languages Linked to (# of outgoing links)
1	Javanese (jav)	Javanese (22,641,844)
		Indonesian (3,761,205)
2	English (eng)	English (11,036,726)
		Javanese (1,443,054)
3	Indonesian (ind)	Indonesian (12,530,636)
		Javanese (4,168,734)
4	Thai (tha)	Thai (4,367,895)
		English (1,775,132)
5	Malay (mly)	Malay (1,944,430)
		Indonesian (1,516,778)
6	Sundanese (sun)	Javanese (2,004,316)
		Sundanese (1,641,623)
7	Luxemburg (ltz)	Luxemburg (1,128,342)
		Javanese (393,524)
8	Occitan (inc)	Occitan (119,734)
		English (83,380)
9	Madurese (mad)	Madurese (387,585)
		Javanese (93,837)
10	Tatar (tat)	Javanese (1,802,601)
		Thai (397,988)

Table 2. Language Link for Indonesian Domain

The table above only shows the top 10 languages. Among these, 8 languages are most frequently linked to the same language. The two other languages are Sundanese and Tatar, both mostly linked to Javanese.

The language graph below shows the languages as the nodes and the edges representing the distance between languages. In the figure above, the six languages of Indonesia are found to be closely connected to each other.

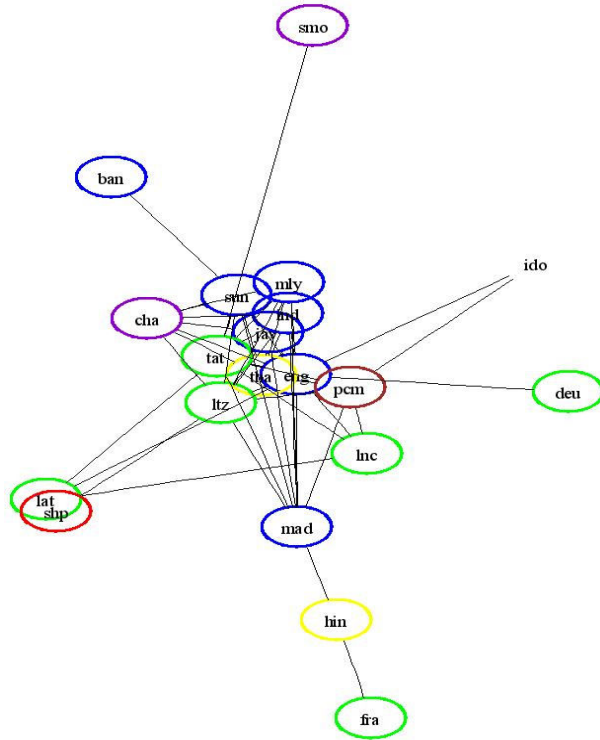


Figure 1. Language Graph of Indonesia

### 5.2 Intermediary Languages

For languages with only a few outgoing links between them, there exists a language that acts as its intermediary, that makes access between the two languages more convenient if passed through.

No.	Intermediary Language	Frequency	Percentage
1	English	89	17.73%
2	Javanese	43	8.57%
3	Tatar	42	8.37%
4	Thai	40	7.97%
5	Madurese	38	7.57%
6	Indonesian	34	6.77%
7	Latin	34	6.77%
8	Sundanese	29	5.78%
9	Malay	28	5.58%
10	Luxemburg	24	4.78%
	Others (top11-20)	122	24.30%
	Total	502	100.00%

Table 3. Intermediary Languages of Indonesia

The above table shows the number of language pairs having the given language as its intermediary language. English has the highest frequency as an

intermediary language. However, it is likely that several pages were misidentified by LIM. The second, Javanese is not surprising since it is a major language of Indonesia. The table below shows selected language pairs, where one of its intermediary languages is Javanese.

Language <i>i</i>	Language <i>j</i>	$P_{ij}$	$P_{ik} * P_{kj}^5$
Tatar	Indonesian	0.01753	0.06085
Tatar	Luxemburg	0.00210	0.01193
Samoan	Balinese	0.00000	0.00052

Table 4. Selected Languages of Indonesia in which it's Intermediary Language is Javanese

### 5.3 Graph Properties

This section discusses the size distribution of the SCCs and the diameter of the Filipino language community in Indonesia.

#### SCC size

The SCCs of a graph are those sets of nodes such that for every node, there is a path to all the other nodes. The size of each SCC refers to the number of nodes it contains. Distribution of the sizes of SCCs gives a good understanding of the graph structure of the web, and has important implications for crawling. If most components have large sizes, only a few nodes are needed as seed URLs (a list of starting pages for a crawler) to be able to reach all the other nodes. If all nodes are members of a single SCC, one URL is enough to crawl all pages.

SCC size	1	2	4	16	19	20	26	45	T <sup>6</sup>
# of SCCs	9	1	3	1	1	1	1	1	18
# of nodes	9	2	12	16	19	20	26	45	149

Table 5. SCC size distribution of the Filipino language community in Indonesia

#### SCC diameter

The maximum distance between any two nodes is the diameter. For each node size, the diameter is

<sup>5</sup> k = Javanese

<sup>6</sup> Total

calculated and plotted in the chart. Their corresponding SCC graph is also shown.

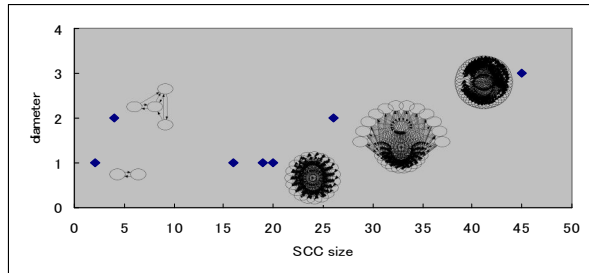


Figure 2. Diameter distribution of the Filipino language community in Indonesia

For the Filipino language subgraph of Indonesia, the component with the largest node size also has the largest diameter. However, the largest diameter size is only 3, which is a very small number. Most of the components have a diameter of 4.

## 6 Implications

For the Filipino language community, many SCCs can be found in the Philippines, since Filipino is one of its national languages. However, for some countries, there are not many SCCs. For example, Indonesia only has 18 SCCs, half of which consist only of one node. However, the largest component size is 45, and there are 4 more large components. By picking out just one node from each SCC with a large size and using it as a seed URL, many web pages can already be downloaded. Add to it the depth parameter of 3, which is the largest diameter, these web pages can be downloaded within a short period of time.

The choice of seed URL and the crawling depth are useful parameters for crawling. The analysis is done for each language community to get these parameters for language-specific crawling purposes. These parameters are different for each language community. This paper shows only shows the case of the Filipino language community of Indonesia as a sample illustration of the diameter metric.

## 7 Conclusion

The vastness and multilingual content of the web makes it a rich source of culturally-diversified information. Since web pages are connected by hy-

perlinks, information can be readily accessed by jumping from one page to another via the hyperlinks.

For each country domain, the web pages written in the same language form a language community. The link structure between language communities shows how connected a language community is with another language community. It can be assumed that the close links between two language communities on the web imply the existence of multilingual speakers of the two languages. Otherwise linked pages will not be visited. In this context, the language graph analysis demonstrated in this study gives an effective tool to understand the linguistic scenes of the country. If the same analysis is performed for the secondary level domain data, further insight into the socio-linguistic status of each language can be drawn. Secondary domain corresponds to different social area of language activities, such as “ac” or “edu” for academic and education arena, “go” or “gov” for government or public arena, and “co” or “com” for commercial business and occupational arena. Although this study does not extend its scope to the secondary level domain analysis, the effectiveness of the approach was demonstrated.

Another implication drawn from this study is that the language graph analysis can identify intermediary languages in the multilingual communities. In the real world, some languages are acting as a medium of communications among the different language speakers. In most cases, such lingua franca are international languages such as English, French, Arabic, etc. But it’s difficult to identify which language is acting as such in detail. But on the web link structure among languages, the language graph can give us a clue to identify this. As shown in this paper, there are a number of languages acting as intermediary between two languages having only a few hyperlinks between them. Although the result of this category is doubtful because of misidentification of language, some cases show the expected result.

The second objective of the study is to give a microscopic level structure of the web communities for much more practical and technical reasons, such as how to design more effective crawling strategy, and how to prepare starting URLs with minimal efforts. The key issue in this context is to reveal the connectedness of the web. To show the connectedness of language communities, several

graph theory metrics, the size and numbers of strongly-connected components and the diameters are calculated and visual presentations of language communities are also given. This information can aid in defining parameters used for crawling, particularly language-specific crawling.

As a summary, the link structure analysis of language graphs can be a useful tool for various spectrums of socio-linguistic and technical research purposes.

## 8 Limitations and Future Work

The results of this research are highly dependent on the language identification module (LIM). With a more improved LIM, more accurate results can be presented. Currently, there is an ongoing experiment that uses a new LIM.

This analysis will also be done to the secondary-level-domains to show the language distribution for different social areas.

Future work also includes the creation of a language-specific crawler that will incorporate the results derived from the analysis of the SCC size and diameter of the language subgraphs.

## Acknowledgment

The study was made possible by the financial support of the Japan Science and Technology Agency (JST) under the RISTEX program and the Asian Language Resource Network Project. We also thank UNESCO for giving official support to the project since its inception.

## References

- Balakrishnan, Hemant and Narsingh Deo. 2006. Evolution in Web graphs. *Proceedings of the 37th South-eastern International Conference on Combinatorics, Graph Theory, and Computing*. Boca Raton, FL.
- Bharat Krishna, Bay-Wei Chang, Monika Henzinger, and Matthias Ruhl. 2001. Who links to whom: mining linkage between Web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51-58, San Jose, California.
- Boldi, Paolo, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. 2004. UbiCrawler: A scalable fully distributed web crawler. *Software: Practice & Experience*, 34(8):711-726.
- Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. In *Proceedings of the 9th World Wide Web Conference*, pages 309-320, Amsterdam, Netherlands.
- Chakrabarti, Soumen, Martin van den Berg, and ByronDom. 1999. Focused Crawling: a new approach to topic-specific Web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*, pages 1623-1640, Toronto, Canada.
- Herring, Susan C., John C. Paolillo, Irene Ramos-Vielba, Inna Kouper, Elijah Wright, Sharon Stoerger, Lois Ann Scheidt, and Benjamin Clark. 2007. Language Networks on LiveJournal. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, Hawaii.
- Kumar, Ravi, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew S. Tompkins, and Eli Upfal. 2000. The Web as a graph. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 1-10, Dallas, Texas, United States.
- Petricek, Vaclav, Tobias Escher, Ingemar J. Cox, and Helen Margetts. 2006. The web structure of e-government: developing a methodology for quantitative evaluation. In *Proceedings of the 15th International World Wide Web Conference*, pages 669-678, Edinburgh, Scotland.
- Pingali, Prasad, Jagadeesh Jagarlamudi, and Vasudeva Varma. 2006. WebKhoj: Indian language IR from Multiple Character Encodings. In *Proceedings of the 15th International World Wide Web Conference*, pages 801-809, Edinburgh, Scotland.
- Stamatakis, Konstantinos, Vangelis Karkaletsis, Georgios Paliouras, James Horlock, Claire Grover, James R. Curran, and Shipra Dingare. 2003. Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pages 75-78. Edinburgh, UK.
- Suzuki, Izumi, Yoshiki Mikami, Ario Ohsato, and Yoshihide Chubachi. 2002. A language and character set determination method based on N-gram statistics. *ACM Transactions on Asian Language Information Processing*, 1(3): 269-278.
- Tamura, Takayuki, Kulwadee Somboonviwat, and Masaru Kitsuregawa. 2007. A method for language-specific Web crawling and its evaluation. *Systems and Computers in Japan*, 38(2):10-20.



# A Multilingual Multimedia Indian Sign Language Dictionary Tool

<b>Tirthankar Dasgupta</b>	<b>Sambit Shukla</b>	<b>Sandeep Kumar</b>	<b>Synny Diwakar</b>	<b>Anupam Basu</b>
IIT, Kharagpur	NIT, Rourkela	NIT, Allahabad.	NIT, Suratkal	IIT, Kharagpur
tirtha@iitkgp	sks.at.nit	mnnit.sand	sunny.diwaka	Anupam-
.ernet.in	r@gmail.co	eep@gmail.	rnitk@gmail.	bas@gmail.
	m	com	com	com

## Abstract

This paper presents a cross platform multi-lingual multimedia Indian Sign Language (ISL) dictionary building tool. ISL is a linguistically under-investigated language with no source of well documented electronic data. Research on ISL linguistics also gets hindered due to a lack of ISL knowledge and the unavailability of any educational tools. Our system can be used to associate signs corresponding to a given text. The current system also facilitates the phonological annotation of Indian signs in the form of HamNoSys structure. The generated HamNoSys string can be given as input to an avatar module to produce an animated sign representation.

## 1 Introduction

A **sign language** is a visual-gesture language that uses hand, arm, body, and face to convey thoughts and meanings. It is a language that is commonly developed in deaf communities, which includes deaf people, their friends and families as well as people who are hard of hearing. Despite common misconceptions, sign languages are complete natural languages, with their own syntax and grammar. However, sign languages are not universal. As is the case in spoken language, every country has got its own sign language with high degree of grammatical variations.

The sign language used in India is commonly known as Indian Sign Language (henceforth called ISL). However, it has been argued that possibly the same SL is used in Nepal, Sri Lanka,

Bangladesh, and border regions of Pakistan (Zeshan et al., 2004). Different dialects of ISL with broad lexical variation are found in different parts of the Indian subcontinent. However, the grammatical structure is same for all dialects (Zeshan, 2003).

The All India Federation of the Deaf estimates around 4 million deaf people and more than 10 million hard of hearing people in India (Zeshan et al, 2004). Studies revealed that, one out of every five deaf people in the world are from India. More than 1 million deaf adults and around 0.5 million deaf children uses Indian Sign Language as a mode of communication (Zeshan et al, 2004). However, an UNESCO report (1980) found that only 5% of the deaf get any education in India. The reason behind such a low literacy rate can be due to the following reasons: a) Till the early 20<sup>th</sup> century, deafness in India, is considered as a punishment for sins and signing is strictly discouraged (Zeshan et. al, 2004). b) Until the late 1970's, it has been believed that, there were no such language called ISL. c) Lack of research in ISL linguistics. d) Unavailability of well documented and annotated ISL lexicon. e) Unavailability of any ISL learning tool. f) Difficulties in getting sign language interpreters.

Linguistic studies on ISL were started around 1978 and it has been found that ISL is a complete natural language, instigated in India, having its own morphology, phonology, syntax, and grammar (Vasishta et. al, 1978; Zeshan et.al, 2004). The research on ISL linguistics and phonological studies get hindered due to lack of linguistically annotated and well documented ISL data. A dictionary of around 1000 signs in four different regional varieties was released (Vasishta et.al,

1978). However, these signs are based on graphical icons which are not only difficult to understand but also lack phonological features like movements and non-manual expressions.

As it has been specified above, ISL is not only used by the deaf people but also by the hearing parents of the deaf children, the hearing children of deaf adults and hearing deaf educators (Zeshan et al, 2004). Therefore the need to build a system that can associate signs to the words of spoken language, and which can further be used to learn ISL, is significant. Further associating signs of different SL (like ASL<sup>1</sup>, BSL<sup>2</sup> and ISL) to a word will help the user to learn foreign SLs simultaneously.

Several works have been done on building multimedia-based foreign SL dictionaries as discussed in (Buttussi et. al., 2007). However no such system is currently available for ISL. moreover, most of the current systems suffer from the following limitations:

- Most of the systems are native language specific and hence, cannot be used for ISL.
- Most of the systems provide a word-sign search but very few systems provide a sign-word or sign-sign search.
- Very few systems are cross platform.
- Systems lack sophisticated phonological information like hand-shape, orientations, movements, and non-manual signs.

In order to overcome the above mentioned crisis, and based on the limitations of the current systems, our objective is to:

- Build a cross platform multilingual multimedia SL-Dictionary tool which can be used to create a large SL lexicon.
- This tool can be used to associate signs to the words, phrases, or sentences of a spoken language text.
- The sign associated with each word is composed of its related part-of-speech and semantic senses.
- The input text (word, phrase, or a sentence) may be in any language (like English or Hindi) and the associated sign can be in any standard sign language (ASL or ISL).
- This tool can also be used to associate complex SL phonological features like hand shape, palm

orientation, locations, movements, and non-manual expressions.

- The phonological features are expressed in terms of HamNoSys (Prillwitz et. al, 1989).
- Facilitate search options like word-sign and search by HamNoSys.
- The generated lexicon is exported in XML file format and the sign is stored in the form of digital videos.
- The video segments are captured using webcams connected with the system. It is possible to attach multiple webcams to the system to capture video segments from multiple angles. This feature enables a user to better understand some of the complex sign language attributes.

The organization of the paper is as follows: Section 2 gives a brief introduction to ISL phonology. Section 3 presents related works on ISL Dictionary. Section 4 presents the overall system architecture of the SL-dictionary tool. Section 5 and 6 presents a brief discussion related HamNoSys representation, and the HamNoSys editor. Section 7 presents conclusion and future work.

## 2 ISL Phonology

Indian Sign Language (ISL) is a visual-spatial language which provides linguistic information using hands, arms, face, and head/body postures. The signer often uses the 3D space around his body to describe an event (Zeshan, 2003). Unlike spoken languages where the communication medium is dependent on sound, in sign language, the communication medium depends upon the visual channel. In spoken language, a word is composed of phonemes. Two words can be distinguished by at least one phoneme. In SL, a sign is composed of cheremes<sup>3</sup> and similarly two signs can differ by at least one chereme (Stokoe, 1978). A sign is a sequential or parallel construction of its manual and non-manual cheremes. A manual chereme can be defined by several parameters like:

- Hand shape.
- Hand location
- Orientation.
- Movements (straight, circular or curved)

<sup>1</sup> ASL: American Sign Language.

<sup>2</sup> BSL: British Sign Language.

<sup>3</sup> The term chereme (originally proposed by William Stokoe (Stokoe, 1978)) in Greek means "hand". It is equivalent to the phonemes of spoken languages.

Non-manual cheremes are defined by:

- Facial expressions.
- Eye gaze and Head/body posture (Zeshan, 2003).

However, there exist some signs which may contain only manual or non-manual components. For example the sign “Yes” is signed by vertical head nod and it has no manual component.

ISL signs can be generally classified into three classes: One handed, two handed, and non-manual signs. Fig. 1 shows the overall Indian sign hierarchy.

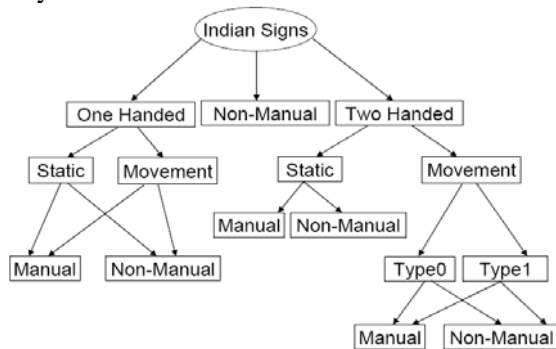


Fig. 1: ISL Type Hierarchy

*One handed signs:* the one handed signs are represented by a single dominating hand. One handed signs can be either static or movement related. Each of the static and movement signs is further classified into manual and non-manual signs. Fig. 2 shows examples of one handed static signs with non-manual and manual components.

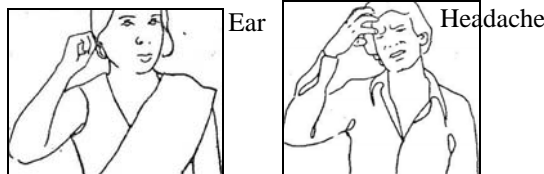


Fig. 2: One Handed static manual sign (Ear) and non-manual sign (Headache).

*Two hand signs:* As in the case of one hand signs, similar classification can be applied to two handed signs. However, two handed signs with movements can be further distinguished as:

*Type0:* Signs where both hands are active (see Fig 3).

*Type1:* Signs where one hand (dominant) is more active compared to the other hand (non-dominant) as shown in Fig 3.

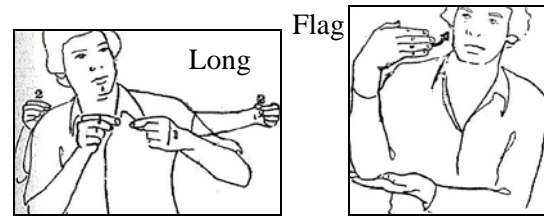


Fig.3 : Two handed sign "long"(both the hands are moving) and “Flag” (only the dominant right hand is moving)

### 3 Related works on ISL dictionary

Linguistic studies on ISL are in their infancy as compared to other natural languages like English, Hindi, or Bengali and also to other SLs. Linguistic work on ISL began during late 1970’s. Before that, the existence of ISL was not acknowledged. In 1977 a survey was conducted (see Vasistha et. al., 1998 for documentation) and it was revealed that ISL is a complete natural language instigated at the Indian subcontinent. Vasistha collected signs from four major states of India (Delhi, Mumbai, Kolkata, and Bangalore) and released four dictionaries of ISL regional varieties. The Ramkrishna Mission vidyalaya, Coimbatore has published another ISL dictionary in 2001. However, all these dictionaries are based on iconic representations of signs. As a result some of the important phonological information like, movements and non-manual expression gets lost. No other work of its kind has so far been reported (Zeshan, 2004).

Several works have been done in building ASL and BSL dictionary tools. Some of the systems are briefly discussed below:

- (Wilcox et. al, 1994) developed a multimedia ASL dictionary tool, which prerecorded digital video frames.
- (Geitz et.al, 1996) developed a VRML based ASL finger spelled system, which ran on internet.
- Sign Smith (VCOM3D, 2004) is a 3D illustrated dictionary of ASL. It is also used as educational software as well as an authoring tool to create ASL content.
- (Buttussi et. al, 2007) proposes an Italian Sign Language dictionary tool. This tool uses H-animator to generate signing avatar. This tool provides multiple search functionality like word-sign, sign-word, and sign-sign search. This tool also facilitates association of one or more SL for a given input word.

## 4 SL-Dictionary

The primary objective of the SL-dictionary tool is to provide an easy to use GUI to create a multilingual multimedia SL dictionary by which a user can associate signs as well as the parameters defining a sign, corresponding to a given text. The overall architecture of the system is shown in Fig. 4. The system has been divided into two modules: a) Expert module and b) User Module.

The expert module has got three main units: a) Input Text Processing Unit b) Visual Data Capture Unit (VDCU) c) Sign Storage Unit and d) HamNoSys Editor.

*Input Text Processing Unit:* In this unit a SL expert chooses the input spoken language (like, English, or Hindi) and the target sign language (like, ISL, or ASL) and then enters a text. The input to the system may be word, phrase, or sentences. If the text is a word the system generates all possible meanings, with the help of WordNet<sup>4</sup>, along with the part of speech (POS)<sup>5</sup> of that particular word. In order to get the exact part-of-speech of a word, the SL expert has to enter an example sentence corresponding to that word. This sentence is given as an input to the POS-tagger to get the correct POS of the word. A word may have multiple senses as returned by WordNet. The user can select one or more senses from the list.

*Visual Data Capture Unit:* Sign corresponding to a word sense is signed by the user which is captured by the Visual Data Capture Unit (VDCU). The VDCU is connected through multiple webcams, placed at different angular positions with respect to the signer. As a result different articulation points of a signs are getting stored with in the database. This will enable the SL learner to understand a particular sign easily. Fig.5 shows how a sign from multiple angles is getting captured.

*Storage Unit:* The input text along with its annotated information, the digital video sign, and the phonological parameters defining the sign are stored with in a database which is further exported into an XML formatted file (see Fig. 6). The phonological parameters are expressed in the form of HamNoSys (discussed in section 5).

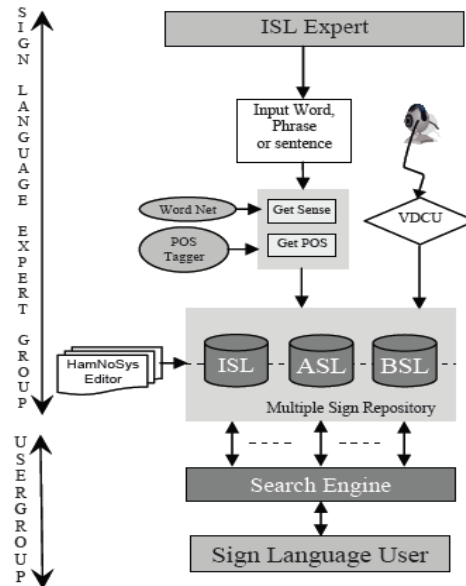


Fig. 4: System Architecture of ISL-Dictionary



Fig.5: capturing video signs from multiple angle

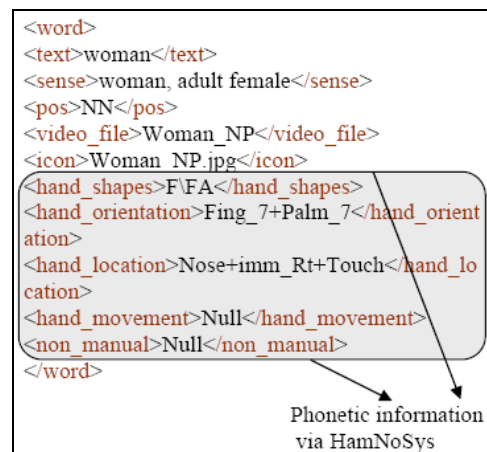


Fig.6: The ISL-dictionary XML Format

*Searching:* The search engine of the current system takes a spoken language text as input parses the XML formatted dictionary and sequentially searches the dictionary. If a match is found, then

<sup>4</sup> [wordnet.princeton.edu/](http://wordnet.princeton.edu/)

<sup>5</sup> We have used the Stanford Part-of-Speech tagger ([nlp.stanford.edu/software/tagger.shtml](http://nlp.stanford.edu/software/tagger.shtml))

the sign corresponding to the lexical entry is being displayed.

### 5 Sign language notation systems

As it has been mentioned above, Sign language does not have any written form. Hence, In order to define a sign we need some notation system. There are a number of phonological notation systems for the representation of SL as discussed in (Smith et.al, 2003). One of the popular among them is Stokoe notation (Stokoe, 2003; Smith et.al, 2003). Stokoe defines a sign by three parameters: a) Hand-shape or designator (*dez*) b) location or place of articulation with respect to the body (*tab*) and c) movements or signation (*sig*).

HamNoSys (Prillwitz et. al, 1989) is a phonetic transcription system, based on Stokoe notation, used to transcribe signing gestures. It is a syntactic representation of a sign to facilitate computer processing. HamNoSys extends the traditional Stokoe based notation system by further expanding sign representation by some more parameters. These parameters can be defined as:

- Dominant hand’s shape.
- Location of the dominant and the non-dominant hand with respect to the body.
- Extended finger orientation of both dominant and non-dominant hand.
- Palm orientation of both hands.
- Movements (straight, circular, or curved)
- Non-manual signs.

Fig. 7 shows examples of different HamNoSys symbols and their descriptions.








	Symbol	Description
	♩	indexfinger stretched
	△	extended finger ahead
	0	palm orientated left
	☐	location shoulder height
	↪	fully stretched out
	↑	hand move ahead
	→	hand move right

Fig.7: HamNoSys symbols and there descriptions

Fig. 8 shows an example where HamNoSys representation of the word “WOMAN” is explained. Here, the parameters like movement and non-

manual signs are not present, as the sign “WOMAN” in ISL does not have these expressions. Fig.9 shows the ISL representation of “WOMAN”.

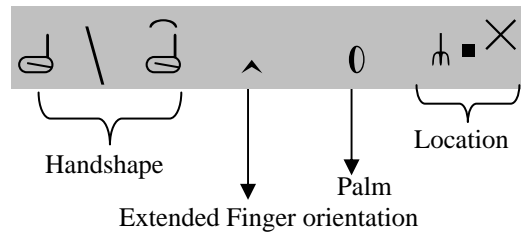


Fig. 8: HamNoSys representation of “WOMAN”



Fig.9: Sign of “WOMAN”

### 6 HamNoSys Editor

Transcribing a sign by HamNoSys is not a trivial task. A user who is transcribing a sign should be an expert in both HamNoSys as well as ISL. Moreover he has to remember all the HamNoSys symbols and their corresponding meanings in order to define a sign. In India it is very difficult to find such a person. Hence our main goal behind building a HamNoSys editor is that, it can be used by an ISL expert with little or no knowledge in HamNoSys. The tool should provide an easy to use GUI that can be used to transcribe phonological information of a sign.

The HamNoSys editor provides a set of graphical images (most of the images are collected from [www.sign-lang.uni-amburg.de/Projekte/HamNoSys](http://www.sign-lang.uni-amburg.de/Projekte/HamNoSys)) for most of the phonological parameters of a sign, like, Hand-shape, orientation, location and movements. Based on the parameters, an ISL expert can choose a set of images and the system will automatically generate the corresponding HamNoSys of the sign. This HamNoSys string can be given as an input to a signing avatar module to generate animated sign representation.

A signing avatar is a virtual human character that performs sign language. However, this character needs a set of instructions which will guide its movement. These instructions can be provided in the form of HamNoSys (Marshall and Sáfár, 2001).

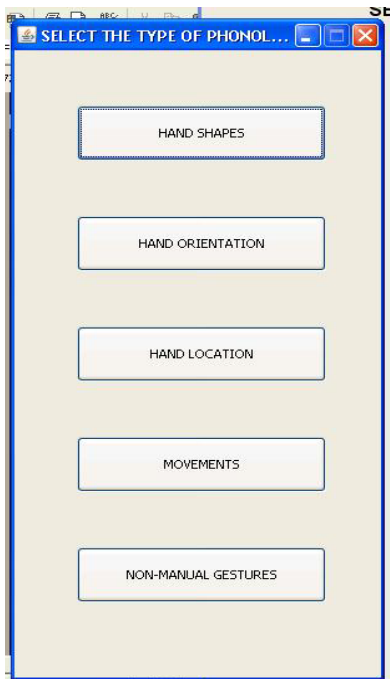


Fig.10: HamNoSys Parameters

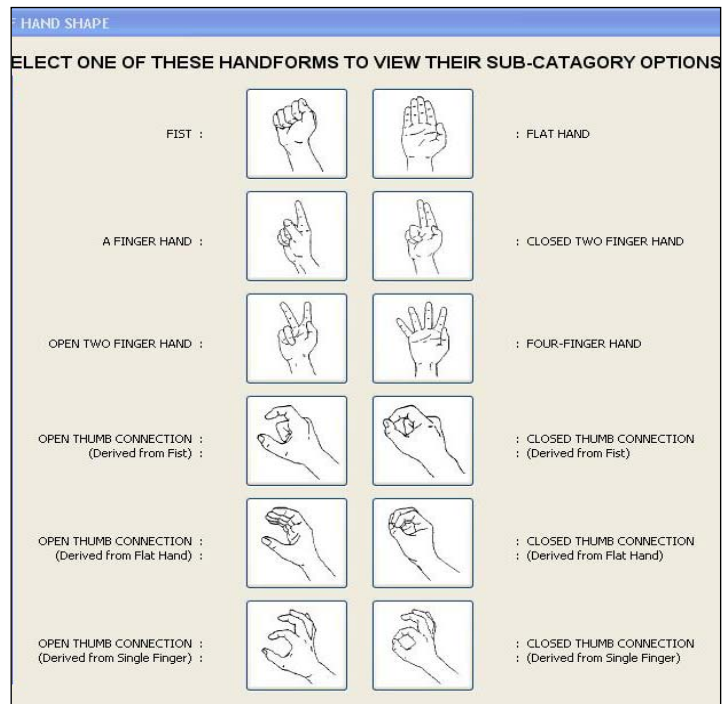


Fig. 11: Twelve basic hand-shape classes

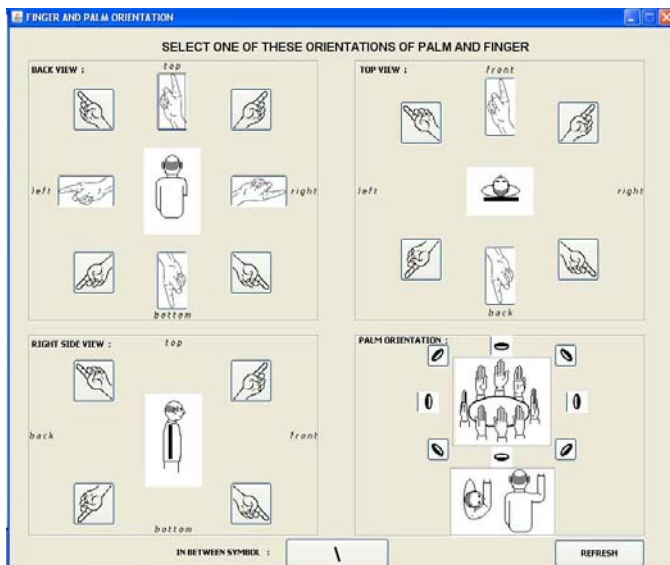


Fig.12: GUI to express finger and palm orientations

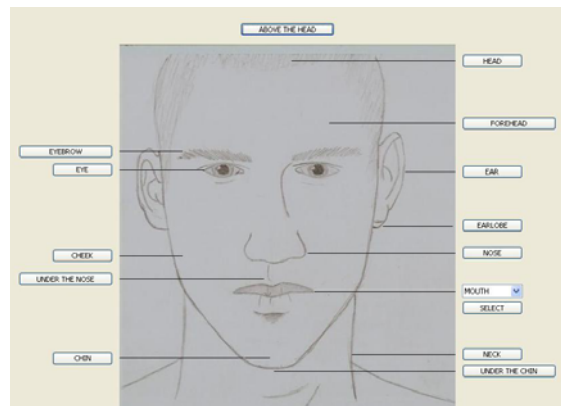


Fig.13: GUI to choose various hand locations near the human face

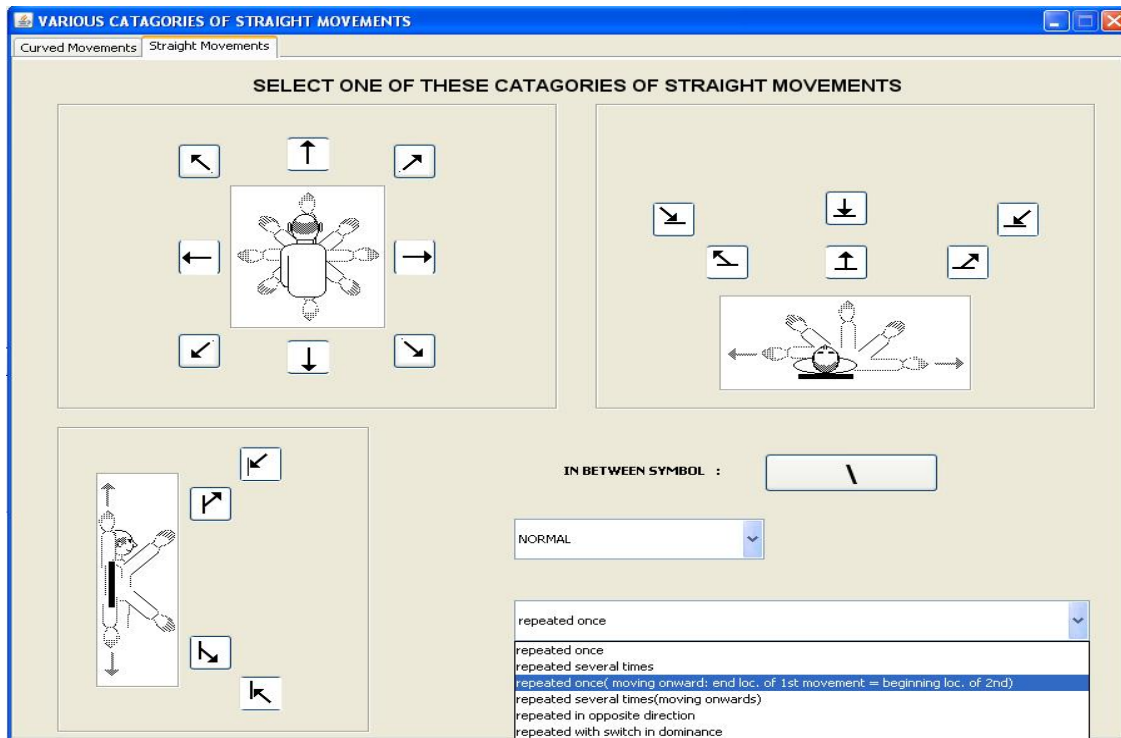


Fig 14: GUI showing various straight movement parameters

Fig.10 shows the five basic parameters of HamNoSys. For each of these parameters there exist interfaces through which a SL expert can choose the desired parameters to define a sign. For example, the right hand side of Fig.11 shows the twelve basic hand-shape classes. Each of these base hand-shapes may contain several derived hand-shapes as defined in HamNoSys (version 4.0). If a particular hand-shape is selected, then the HamNoSys symbol corresponding to the hand-shape gets stored in the XML database (see Fig.6). Similarly, separate interfaces have been provided to identify palm orientation (see Fig.12), hand location (see Fig.13), movements (see Fig.14), and non-manual signs.

Due to its symbolic structure, HamNoSys is fairly easy to write, and understand. However, there are some drawbacks on this notation system that make it difficult to be used universally for all sign languages (Smith and Edmondson, 2004). For example, HamNoSys uses some fixed set of symbols to define a sign however it is possible that a particular sign in any sign language may not be defined by 'the fixed set of symbols. For example HamNoSys does not have well defined symbols for non-manual expressions. Consider the sign

“BITTER”, in ISL the representation is shown in Fig.15. It can be observed that it is very difficult to represent the facial expressions like eyebrow by HamNoSys. Currently we have a collection of around 979 sign icons (published by Vasistha et. al 1998), which we are trying to transcribe in HamNoSys. Out of these, 16% of the signs contain non-manual features which we are unable to represent in HamNoSys.



Fig.15: ISL representation of "BITTER"

## 7 Conclusion and Future works

The paper presents an approach towards building a multimedia SL dictionary tool. This tool can be used to prepare a well documented ISL dictionary. The system is intended to take any Indian language text as input and can store signs in any SL. Currently the system takes English, Hindi and Bengali texts as input and can store signs in ISL only. The system also provides an easy to use GUI

to include phonological information of a sign in the form of HamNoSys string. The generated HamNoSys string can then be used as an input to the signing avatar module to produce animated sign output.

In the next phase of our work we will improve the system so that it can associate signs in any other SL (like, ASL and BSL). Further, WordNet as well as POS Tagger corresponding to Hindi and Bengali languages should also be integrated with the system. Also, support has to be built so that system can perform sign-to-word and sign to sign search. We will also perform proper evaluation of the HamNoSys editor in order to understand its utility to the SL user.

## References

- Buttussi F., Chittaro L., Coppo M. 2007. Using Web3D technologies for visualization and search of signs in an international sign language dictionary. Proceedings of the twelfth international conference on 3D web technology. Perugia, Italy Pages: 61 – 70 Year of Publication: 2007 ISBN:978-1-59593-652-3
- Geitz, S., Hanson, T., Maher, S. 1996. Computer generated 3-dimensional models of manual alphabet hand-shapes for the World Wide Web. In *Assets '96: Proceedings of the second annual ACM conference on Assistive technologies*, ACM Press, New York, NY, USA, 27–31.
- Marshall I. and Sáfár É. 2001.Extraction of semantic representations from syntactic SMU link grammar linkages.. In G. Angelova, editor, Proceedings of Recent Advances in Natural Lanugage Processing, pp: 154-159, Tzigov Chark, Bulgaria, September.
- Prillwitz P., Regina Leven, Heiko Zienert, Thomas Hamke, and Jan Henning. 1989. HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide, volume 5 of International Studies on Sign Language and Communication of the Deaf. Signum Press, Hamburg, Germany,
- Smith G., Angus. 1999. English to American Sign Language machine translation of weather reports. Proceedings of the Second High Desert Student Conference in Linguistics.
- Smith,K.C., Edmondson, W. 2004. The Development of a Computational Notation for Synthesis of Sign and Gesture, *GW03*(312-323).
- Speers, A. 1995. SL-Corpus: A computer tool for sign language corpora., Georgetown University.
- Stokoe W. C., 1960. Sign language structure: an outline of the visual communication systems of the American deaf. 2nd edition, 1978. Silver Spring, MD: Linstok Press.
- VCOM3D,2004. Sign smith products. <http://www.vcom3d.com>.
- Wilcox, S., Scheibman, J., Wood, D., Cokely, D., and stokoe, w. c. 1994. Multimedia dictionary of American Sign Language. In *Assets '94: Proceedings of the first annual ACM conference on Assistive technologies*, ACM Press, New York, NY, USA, 9–16.
- Vasishta M., Woodward J., DeSantis S. 1998, “An Introduction to Indian Sign Language”, All India Federation of the Deaf (Third Edition).
- Zeshan U., 2003,”Indo-Pakistani Sign Language Grammar: A Typological Outline”, *Sign Language Studies - Volume 3, Number 2, , pp. 157-212*
- Zeshan U., Madan M. Vasishta, Sethna M. 2004, “implementation of indian sign language in educational settings”- *Volume 15, Number 2, Asia Pacific Disability Rehabilitation Journal, , pp. 15-35*



# A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus

**Deniz Zeyrek**

Department of Foreign Language  
Education  
Middle East Technical University  
Ankara, Turkey  
dezeyrek@metu.edu.tr

**Bonnie Webber**

School of Informatics  
University of Edinburgh  
Edinburgh, Scotland

bonnie@inf.ed.ac.uk

## Abstract

This paper describes first steps towards extending the METU Turkish Corpus from a sentence-level language resource to a discourse-level resource by annotating its discourse connectives and their arguments. The project is based on the same principles as the Penn Discourse TreeBank (<http://www.seas.upenn.edu/~pdtb>) and is supported by TUBITAK, The Scientific and Technological Research Council of Turkey. We first present the goals of the project and the METU Turkish corpus. We then describe how we decided what to take as explicit discourse connectives and the range of syntactic classes they come from. With representative examples of each class, we examine explicit connectives, their linear ordering, and types of syntactic units that can serve as their arguments. We then touch upon connectives with respect to free word order in Turkish and punctuation, as well as the important issue of how much material is needed to specify an argument. We close with a brief discussion of current plans.

## 1 Introduction

The goal of the project is to extend the METU Turkish Corpus (Say et al, 2002) from a sentence-level language resource to a discourse-level resource by annotating its discourse connectives,

and their arguments. The 2-million word METU Turkish Corpus (MTC) is an electronic resource of 520 samples of continuous text from 291 different sources written between 1990-2000. It includes multiple genres, such as novels, short stories, newspaper columns, biographies, memoirs, etc. annotated topographically, i.e., for paragraph boundaries, author, publication date, and the source of the text. A small part of the MTC, called the METU-Sabancı TreeBank (5600 sentences) has been annotated with morphological features and dependency relationships (e.g., modifier-of, subject-of, object-of, etc.). The result is a set of dependency trees. The MTC as a whole provides a large-scale resource on Turkish discourse and is being used in research on Turkish. To date, there have been 81 requests for permission to use the MTC and 31 requests to use the TreeBank sub-corpus. Most of the users are linguists, computer or cognitive scientists working on Turkish, or graduate students of similar disciplines. Some users have expressed a desire for the MTC to be extended by annotations at the discourse level, which provides further impetus for the present project.

The result of annotating discourse connectives will be a clearly defined level of discourse structure on the MTC. Annotation of text from the multiple genres present in the MTC will allow us to compare the distribution of connectives and their arguments across genres. The annotation will help researchers understand Turkish discourse by enabling them to give concise, clear descriptions of the issues concerning discourse structure and semantics, and support a rigorous empirical

characterization of where and how the free word-order in a language like Turkish is sensitive to features of the surrounding discourse. It can thus serve as a major resource for natural language processing, language technology and pedagogy.

## 2 Overview of Turkish Discourse Connectives

From a semantic perspective, a discourse connective is a predicate that takes as its arguments, abstract objects (propositions, facts, events, descriptions, situations, and eventualities). The primary linguistic unit in which abstract objects (AOs) are realized in Turkish is the clause, either tensed or untensed. Discourse connectives themselves may be realized explicitly or implicitly. An explicit connective is realized in the form of a lexical item or a group of lexical items, while an implicit connective can be inferred from adjacent text spans that realise AOs and whose AOs are taken to be related. To constrain the amount of text selected for arguments, a *minimality principle* can be imposed, limiting arguments to the minimum amount of information needed to complete the interpretation of the discourse relation. The project will initially focus on annotating explicit connectives, integrating implicit ones at a later stage.

One of the most challenging issues so far has been determining the set of explicit discourse connectives in Turkish (i.e., the various linguistic elements that can be interpreted as predicates on AO arguments) and the syntactic classes they are identified with. In the Penn Discourse TreeBank (PDTB), the explicit discourse connectives were taken to comprise (1) coordinating conjunctions, (2) subordinating conjunctions, and (3) discourse adverbials (Forbes-Riley et al., 2006). But coordinating and subordinating conjunctions are not classes in Turkish *per se*. Moreover, most of the existing grammars of Turkish describe clausal adjuncts and adverbs in semantic (e.g., temporal, additive, resultative, etc.) rather than syntactic terms. We therefore made a rough classification first and determined the broad syntactic classes by considering the morpho-syntactic properties shared by elements of the initial classification.

As a result of this process, we have come to identify explicit discourse connectives in Turkish with three grammatical types, forming five classes:

- (a) Coordinating conjunctions such as single lexical items *çünkü* ‘because’, *ama* ‘but’, *ve* ‘and’, and the particle *da*. (N.B., *da* can also function as a subordinator.)
- (b) Paired coordinating conjunctions such as *hem .. hem* ‘both and’, *ne .. ne* ‘neither nor’ which link two clauses, with one element of the pair associated with each clause in the discourse relation.
- (c) Simplex subordinators (also termed as converbs), i.e., suffixes forming non-finite adverbial clauses, e.g. *-(y)kAn*, ‘while’, *-(y)ArAk* ‘by means of’.
- (d) Complex subordinators, i.e., connectives which have two parts, usually a postposition (*rağmen* ‘despite’, *için* ‘for’, *gibi* ‘as well as’) and an accompanying suffix on the (non-finite) verb of the subordinate clause.<sup>1</sup>
- (e) Anaphoric connectives such as *ne var ki* ‘however’, *üstelik* ‘what is more’, *ayrıca* ‘apart from this’, *ilk olarak* ‘firstly’, etc.

In the PDTB, non-finite clauses have not been annotated as arguments. However, since all non-finite clauses are marked with a suffix in Turkish (see sections 4.1 and 4.2 below) and encode a relation between AOs, we would have missed an important property of the language if we had not identified them as discourse connectives (cf. Prasad et al., 2008).

All the discourse connectives above have exactly two arguments. So as in English, while verbs in Turkish can vary in the number of arguments they take, Turkish discourse connectives take two and only two arguments. These can conveniently be called ARG1 and ARG2. It remains an open question whether there is any language in which discourse connectives take more than two arguments.

In the following, we give representative examples of each of the above five classes of discourse connectives and discuss the assignment of the argument labels, linear order of arguments and types of arguments. By convention, we label

<sup>1</sup> Postpositions correspond to prepositions in English, though there are many fewer of them. They form a subordinate clause by nominalizing their complements and marking them with the dative, ablative, or the possessive case. In the examples given in this paper, suffixes are shown in upper-case letters. Case suffixes are underlined in addition to being presented in upper-case letters.

the argument containing (or with an affinity for) the connective as ARG2 (presented in boldface) and the other argument as ARG1 (presented in italics). Discourse connectives are underlined. This annotation convention is used in the English translations as well. Except for examples (12), (13), (19), (20), all examples have been taken from the MTC.

### 3 Coordinating conjunctions

#### 3.1 Simple coordinating conjunctions

Coordinating conjunctions are like English and combine two clauses of the same syntactic type, e.g., two main clauses. They are typically sentence-medial and show an affinity with the second clause (evidenced in part through punctuation and their ability to move to the end of the second clause). Whether a coordinating conjunction links clauses within a single sentence or clauses across adjacent sentences (cf. Section 6), it shows an affinity with the second clause. Thus ARG2 of these conjunctions is the second clause and ARG1 is the first clause.

- (1) *Yapılarını kerpiçten yapıyorlar, ama sonra taşı kullanmayı öğreniyorlar. Mimarlık açısından çok önemli, **cünkü bu yapı malzemesini başka bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz.***  
*'They constructed their buildings first from mud-bricks but then they learnt to use the stone. Architecturally, this is very important **because we see the use of this construction material with another one at this site for the first time.***

The particle *dA* can serve a discourse connective function with an additive (Example 2) or adversative sense (Example 3). In contrast with coordinating conjunctions, the order of arguments to *dA* is normally ARG2-ARG1, thus exhibiting a similarity with subordinators (see below). However, since *dA* combines two clauses of the same syntactic type, we take it to be a simple coordinating conjunction.

- (2) **Konuşmayı unuttum diyorum da güliyorlar bana.**  
*'I said I've forgotten to talk and they laughed at me.'*
- (3) **Belki bir çocuğumuz olsa onunla oyalanırdım da Allah kısmet etmedi.**

**'If we had a child I would keep myself busy with her/him but God did not predestine it.'**

#### 3.2 Paired coordinating conjunctions

Paired coordinating conjunctions are composed of two lexical items, with the second often a duplicate of the first element. These lexical items express a single discourse relation, such as disjunction as in example (4). The order of arguments is ARG1-ARG2 and the position of the conjunctions is clause-initial.

- (4) *Birilerinin ya işi vardır, aceleyle yürürler, ya koşarlar.*  
*'Some people are either busy and walk hurriedly, or they run.'*

### 4 Subordinators

#### 4.1 Simplex subordinators

When a subordinate clause is reduced in Turkish, it loses its tense, aspect and mood properties. In this way, it becomes a nominal or adverbial clause associated with the matrix verb. The relationship of an adverbial clause with the AO expressed by the matrix verb and its arguments is conveyed by a small set of suffixes corresponding to English 'while', 'when', 'by means of', 'as if', or temporal 'since', added to the non-finite verb of the reduced clause. This pair of non-finite verb and suffix, we call a 'converb'. The normal order of the arguments of a converb is ARG2-ARG1, where the converb appears as the last element of ARG2. The following example illustrates *-(y)ArAk* 'by means of' and its arguments:

- (5) *Kafiye Hanım beni kucakladı, **yanagını yanığıma sürterek** iyi yolculuklar diledi.*  
*'Kafiye hugged me and **by rubbing her cheek against mine**, she wished me a good trip.'*

#### 4.2 Complex subordinators

Complex subordinators constitute a larger set than the set of simplex subordinators. Here, a lexical item, usually a postposition, must appear with a nominalizing suffix and, if required, a case suffix as well. If the verb of the clause does not have a subject, it is nominalized with *-mAk* (the infinitive suffix). If it has a subject, it is nominalized with *-DIK* (past) or *-mA* (non-past) and carries the possessive marker agreeing with the subject of the

verb. The normal order of the arguments of a complex subordinator is the same as with converbs, i.e., ARG2-ARG1. The nominalizer, the possessive and the case suffix (if any) appear attached to the non-finite verb of ARG2 in that order. The connective appears as the last element of ARG2.

Some postpositions have multiple senses, depending on the type of nominalizer attached to the non-finite verb. For example, the postposition *için* means causal ‘since’ with *-DIK* (Example 6), and ‘so as to’ with *-mA* or *-mAk* (Example 7). In these examples, the lexical part of the complex subordinator is underlined, and the suffixes on the non-finite verb of ARG2 rendered in small caps.

- (6) **Herkes çoktan pazara çıktığı için** *kentin o dar, eğri büğrü arka sokaklarını boşalmış ve sessiz bulurduk.*  
‘Since everyone has gone to the bazaar long time ago, we would find the narrow and curved back streets of the town empty and quiet.’
- (7) **[Turhan Baytop] Paris Eczacılık Fakültesi Farmakognozi kürsüsünde görgü ve bilgisini arttırmak için çalışmıştır.**  
‘Turhan Baytop worked at Paris Pharmacology Faculty so as to increase his experience and knowledge.’

Since postpositions also have a non-discourse role in which they signal a verb’s arguments and/or adjuncts, we will only annotate postpositions as discourse connectives when they have clausal elements as arguments. Given that a clausal element always has a nominalizing suffix, the distinction will be straightforward. For example, in (8) *için* takes an NP complement (marked with the possessive case) and will not be annotated, while in (9) *rağmen* ‘despite’ comes with a nominalizer and the dative suffix, and it will be annotated:

- (8) Bunun için paraya ihtiyacımız var.  
‘We need money for this.’
- (9) **Çok iyi bir biçimde yayılmış olmasına rağmen** *Celtis (çitlenbik) polenin yokluğu dikkate değerdir.*  
‘Despite not dispersing well, the absence of the *Celtis [tree] polen* is worthy of attention.’

In general, both parts of a complex subordinator must be realized in the discourse. An exception is ‘if’ *eğer* and its accompanying suffix *-sE* (and the

marker agreeing with the subject of the subordinate clause where necessary). The suffix suffices to introduce a discourse relation on its own, even without the postposition *eğer*:

- (10) **Salman Rushdi öldürülürse** *İslam dini bundan bir onur mu kazanacak?*  
‘If Salman Rushdi was to be killed, would the Islam religion be honoured?’
- (11) **Eğer sigarayı bırakmak için mükemmel zamanı bekliyorsanız** *asla sigarayı bırakamazsınız.*  
‘If you are waiting for the best time to stop smoking, you can never stop smoking.’

## 5 Anaphoric connectives

The fifth type of explicit discourse connectives are anaphoric connectives. Anaphoric connectives are distinguished from clausal adverbs like *çoğunlukla* ‘usually’, *mutlaka* ‘definitely’, *maalesef* ‘regrettably’, which are interpreted only with respect to their matrix sentence. In contrast, anaphoric connectives also require an AO from a sentence or group of sentences adjacent (Example 12) or non-adjacent (Example 13) to the sentence containing the connective. Another important property of anaphoric connectives is that they can access the inferences in the prior discourse (Webber et al 2003). This material is neither accessible by other types of discourse connectives nor clausal connectives. For example, in example (14), the anaphoric connective *yoksa* ‘or else, otherwise’ accesses the inference that the organizations have not united and hence did not introduce political strategies unique to Turkey.

- (12) *Ali hiç spor yapmaz. Sonuç olarak çok istediği halde kilo veremiyor.*  
‘*Ali never exercises. Consequently, he can’t lose weight* although he wants to very much.’
- (13) *Zeynep önceleri Bodrum’da oturdu. Krediyle deniz kenarında bir ev aldı. Evi dayadı, döşedi, bahçeye yasemin ekti. Ne var ki banka kredisini ödeyemediğinden evi satmak zorunda kaldı.*  
‘*Zeynep first lived in Mersin. She bought a house by the sea on credit. She furnished it fully and planted jasmine in the garden. However, she had to sell the house because she couldn’t pay back the credit.*’

- (14) *Bu örgütlerin birleşerek Türkiye'yi etkilemesi ve Türkiye'ye özgü politikaları gündeme getirmesi lazım. Yoksa Tony Blair şöyle yaptı şimdi biz de şimdi böyle yapacağımızla olmaz.*  
'These organizations must unite, have an impact on Turkey and introduce political strategies unique to Turkey. Or else talking about what Tony Blair did and hoping to do what he did is outright wrong.'

## 6 Ordering flexibility of explicit discourse connectives and their arguments

In Turkish, the linear ordering of coordinating conjunctions and subordinators and the clauses in which they occur shows some flexibility as to where in the clause they appear or as to the ordering of the clauses. For example, coordinating conjunctions may appear at the beginning of their ARG2, i.e. S-initially. This was shown earlier in Example (1). The sentences below illustrate *ama* 'but' and *çünkü* 'because' used at this position.

- (15) *Hatem Ağa'nın malına kimse yanaşamaz, dokunamazdı. Ama Osman gitmiş, Hatem Ağa'nın çiftliğini yakmıştı.*  
'No one could approach and touch Agha Hatem's property. But Osman had burnt Agha Hatem's ranch.'
- (16) *Söz özgürlüğünün belli yasalar, belli ilkeler çerçevesinde kalmak zorunda olduğunu biliyoruz. Çünkü, bütün özgürlükler gibi, belli sınırlar aşılnca, başkalarına zarar vermek, başkalarının özgürlüklerini zedelemek söz konusu oluyor.*  
'We know that *freedom of speech should remain within the limits of certain laws and principles.* Because, like all the other freedoms, when certain constraints are violated, one may harm others' freedom.'

But coordinating conjunctions may also appear at the end of their ARG2 and so will appear S-finally in sentences with ARG1-ARG2 order. Below, we illustrate two cases of *ama* 'but' and *çünkü* 'because'.

- (17) *Kazıyabildiğini sildi, biriktirdi mendilinin içine. Çaba isteyen zor bir işti bu yaptığı ama.*  
'He wiped the area he had scraped and saved all he could scrape in his rag. But what he was doing was a difficult job, requiring effort.'
- (18) *Kimi müşteriler dore rengi kumaşlarla, sarı taftalarla gelirdi de, elim dolu yapamam, diye*

*geri çevirirdi, pek anlam veremezdim. Parayı severdi çünkü.*  
'Some customers would come with gold coloured fabrics and yellow taffeta weaves but he would reject them saying his hands were full, which I could not give any meaning to. Because he loved money.'

In contrast, the position of a subordinator (both simplex and complex) in its ARG2 clause is fixed: it must appear at the end of the clause, as shown in example (19). However, the clause is free in the sentence and may be moved to the right of the sentence, as in example (20). It is a matter of empirical research to find out whether different genres vary more in how clauses are ordered and what motivates preposing of ARG1.

- (19) *Ayşe konuşurken ben dinlemiyordum.*  
'I was not listening while Ayşe was talking.'
- (20) *Ben dinlemiyordum Ayşe konuşurken.*  
'I was not listening while Ayşe was talking.'

## 7 Issues and plans

As mentioned above, we also plan to annotate implicit connectives between adjacent sentences or clauses whose relation is not explicitly marked with a discourse connective. This we will do at a later stage, after explicit connectives have been annotated, following the procedure used in annotating implicit connectives in the PDTB (PDTB-Group, 2006). Preliminary analysis has shown that punctuation serves as a useful hint in inserting a coordinating conjunctions such as 'and' or an anaphoric connective such as 'then' or 'consequently' between the multiple adjacent main clauses that can occur in a Turkish sentence separated by a comma. Example (21) illustrates these cases.

- (21) *Yürüyor, Imp = THEN oturuyor, resim yapmaya çalışıyor ama yapamıyor, tabela yazmaya çalışıyor ama yazamıyor, Imp= CONSEQUENTLY sıkılıp sokağa çıkıyor, Imp=AND bisikletine atladığı gibi pedallara basıyor.*  
'He walks around, then sits down and tries to draw, but he can't. He tries to inscribe words on the wooden plaque, but again he can't. Consequently he gets bored, goes out, and hops on his bike and pedals.'

A second important issue that will have to be tackled in the project is determining how much material is needed to specify the argument of a discourse connective. Annotation will be on text spans, rather than on syntactic structure. This reflects two facts: First, there is only a small amount of syntactically treebanked data in the MTC, and secondly, as has been discovered for English, one can not assume that discourse units map directly to syntactic units (Dinesh et al, 2005). Preliminary analysis also shows that discourse units may not coincide with a clause in its entirety. For example, in examples (9) and (16), one can take ARG1 to cover only the nominal complement of the matrix verb: The rest of the clause is not necessary to the discourse relation. The ways in which the arguments of a discourse connective may diverge from syntactic units must be characterized for Turkish as is being done for English (Dinesh et al, 2005).

A third issue we will investigate is whether different senses of a subordinator may be identified simply from the type of nominalizing suffix required on the subordinate verb. For example, we have noted in examples (6) and (7) that the two senses of the postposition için (namely, ‘since (causal)’ and ‘in order to’) are disambiguated by the nominalizing suffixes. The extent to which morphology aids sense disambiguation is an empirical issue that will be further addressed in the project.

## Acknowledgement

We would like to thank Sumru Özsoy, Aslı Göksel and Cem Bozşahin for their comments on an earlier version of this paper. The first author also thanks the Caledonian Research Foundation and the Royal Society of Edinburgh for awarding her with the European Visiting Research Fellowship, which made this research possible. All remaining errors are ours.

## References

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2005). Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, Michigan. June 2005.

Katherine Forbes-Riley, Bonnie Webber and Aravind Joshi (2006). Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics* 23, pp. 55–106.

Aslı Göksel and Celia Kerslake (2005). *Turkish: A Comprehensive Grammar*. London and New York: Routledge.

Kornfilt, Jacklin (1997). *Turkish*. London and New York: Routledge.

PDTB-Group (2006). The Penn Discourse TreeBank 1.0 Annotation Manual. *Technical Report IRCS 06-01*, University of Pennsylvania.

Rashmi Prasad, Samar Husain, Dipti Sharma and Aravind Joshi (2008). Towards an Annotated Corpus of Discourse Relations in Hindi. *The Third International Joint Conference on Natural Language Processing*, January 7-12, 2008.

Bilge Say, Deniz Zeyrek, Kemal Oflazer and Umüt Özge (2002). Development of a Corpus and a TreeBank for Present-day Written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics*, Eastern Mediterranean University, Cyprus, August 2002.

Bonnie Webber, Aravind Joshi, Matthew Stone and Alistair Knott (2003). Anaphora and Discourse Structure. *Computational Linguistics* 29 (4) 547-588.

**Appendix:** A preliminary list of explicit discourse connectives found in the MTC belonging to five syntactic classes and their English equivalents

Simple coordinating conjunctions	English equivalent
ama	but
fakat	but
çünkü	because
dA	and, but
halbuki	despite
oysa	despite
önce	before
sonra	after
ve	and
veya	or
ya da	or
veyahut	or

<b>Paired coordinating conjunctions</b>	<b>English equivalent</b>
hem .. hem	both and
ya .. ya	either or
gerek .. gerek(se)	either or

<b>Simplex subordinators (Converbs)</b>	<b>English equivalent</b>
-ArAk	by means of
-Ip	and
-(y)kEn	while, whereas
-(y)AlI	since
-(I)ncA	when

<b>Complex subordinators</b>	<b>English equivalent</b>
-Ir gibi	as if, as though
-eğer (y)sE	if
-dİğI zaman	when
-dİğI kadar	as much as
-dİğI gibi	as well as
-dAn sonra	after
-dAn önce	before
-dAn dolayı	due to
-(y)sE dA	even though
-(y)İncaya kadar/dek	until
-(y)AlI beri	since (temporal)
-(n)A rağmen/karşılık	despite, although
-(n)A göre	since (causal)

<b>Anaphoric connectives</b>	<b>English equivalent</b>
aksi halde	if not, otherwise
aksine	on the contrary
bu nedenle	for this reason
buna rağmen/karşılık	despite this
bundan başka	besides this
bunun yerine	instead of this
dahası	moreover, in addition
ilk olarak	firstly, first of all
örneğin	for example
mesela	for example
sonuç olarak	consequently
üstelik	what is more
yoksa	otherwise
ardından	afterwards





# Towards an Annotated Corpus of Discourse Relations in Hindi

Rashmi Prasad\*, Samar Husain†, Dipti Mishra Sharma† and Aravind Joshi\*

## Abstract

We describe our initial efforts towards developing a large-scale corpus of Hindi texts annotated with discourse relations. Adopting the lexically grounded approach of the Penn Discourse Treebank (PDTB), we present a preliminary analysis of discourse connectives in a small corpus. We describe how discourse connectives are represented in the sentence-level dependency annotation in Hindi, and discuss how the discourse annotation can enrich this level for research and applications. The ultimate goal of our work is to build a Hindi Discourse Relation Bank along the lines of the PDTB. Our work will also contribute to the cross-linguistic understanding of discourse connectives.

## 1 Introduction

An increasing interest in human language technologies such as textual summarization, question answering, natural language generation has recently led to the development of several discourse annotation projects aimed at creating large scale resources for natural language processing. One of these projects is the Penn Discourse Treebank (PDTB Group, 2006),<sup>1</sup> whose goal is to annotate the discourse relations holding between eventualities described in a text, for example causal and contrastive relations. The PDTB is unique in using a lexically grounded approach for annotation: discourse relations are anchored in lexical items (called “explicit discourse connectives”) whenever they are

explicitly realized in the text. For example, in (1), the causal relation between ‘the federal government suspending US savings bonds sales’ and ‘Congress not lifting the ceiling on government debt’ is expressed with the explicit connective ‘because’.<sup>2</sup> The two arguments of each connective are also annotated, and the annotations of both connectives and their arguments are recorded in terms of their text span offsets.<sup>3</sup>

(1) *The federal government suspended sales of U.S. savings bonds because Congress hasn’t lifted the ceiling on government debt.*

One of the questions that arises is how the PDTB style annotation can be carried over to languages other than English. It may prove to be a challenge cross-linguistically, as the guidelines and methodology appropriate for English may not apply as well or directly to other languages, especially when they differ greatly in syntax and morphology. To date, cross-linguistic investigations of connectives in this direction have been carried out for Chinese (Xue, 2005) and Turkish (Deniz and Webber, 2008). This paper explores discourse relation annotation in Hindi, a language with rich morphology and free word order. We describe our study of “explicit connectives” in a small corpus of Hindi texts, discussing them from two perspectives. First, we consider the type and distribution of Hindi connectives, proposing to annotate a wider range

<sup>2</sup> The PDTB also annotates implicit discourse relations, but only locally, between adjacent sentences. Annotation here consists of providing connectives (called “implicit discourse connectives”) to express the inferred relation. Implicit connectives are beyond the scope of this paper, but will be taken up in future work.

<sup>3</sup> The PDTB also records the senses of the connectives, and each connective and its arguments are also marked for their attribution. Sense annotation and attribution annotation are not discussed in this paper. We will, of course, pursue these aspects in our future work concerning the building of a Hindi Discourse Relation Bank.

\* University of Pennsylvania, Philadelphia, PA, USA, {rjprasad,joshi}@seas.upenn.edu

† Language Technologies Research Centre, IIIT, Hyderabad, India, samar@research.iiit.ac.in, dipti@iiit.ac.in

<sup>1</sup> <http://www.seas.upenn.edu/pdtb>

of connectives than the PDTB. Second, we consider how the connectives are represented in the Hindi sentence-level dependency annotation, in particular discussing how the discourse annotation can enrich the sentence-level structures. We also briefly discuss issues involved in aligning the discourse and sentence-level annotations.

Section 2 provides a brief description of Hindi word order and morphology. In Section 3, we present our study of the explicit connectives identified in our texts, discussing them in light of the PDTB. Section 4 describes how connectives are represented in the sentence-level dependency annotation in Hindi. Finally, Section 5 concludes with a summary and future work.

## 2 Brief Overview of Hindi Syntax and Morphology

Hindi is a free word order language with SOV as the default order. This can be seen in (2), where (2a) shows the constituents in the default order, and the remaining examples show some of the word order variants of (2a).

- (2) a. मलय ने समीर को किताब दी ।  
malay ERG sameer DAT book gave  
“Malay gave the book to Sameer” (S-IO-DO-V)<sup>4</sup>  
b. मलय ने किताब समीर को दी. (S-DO-IO-V)  
c. समीर को मलय ने किताब दी. (IO-S-DO-V)  
d. समीर को किताब मलय ने दी. (IO-DO-S-V)  
e. किताब मलय ने समीर को दी. (DO-S-IO-V)  
f. किताब समीर को मलय ने दी. (DO-IO-S-V)

Hindi also has a rich case marking system, although case marking is not obligatory. For example, in (2), while the subject and indirect object are explicitly for the ergative (ERG) and dative (DAT) cases, the direct object is unmarked for the accusative.

## 3 Discourse Connectives in Hindi

Given the lexically grounded approach adopted for discourse annotation, the first question that arises is how to identify discourse connectives in Hindi. Unlike the case of the English connectives in the PDTB, there are no resources that alone or together provide an exhaustive list of connectives in the

language. We did try to create a list from our own knowledge of the language and grammar, and also by translating the list of English connectives in the PDTB. However, when we started looking at real data, this list proved to be incomplete. For example, we discovered that the form of the complementizer ‘कि’ also functions as a temporal subordinator, as in (3).

- (3) [ वह बाल्टी के गंदे पानी से अपनी चॉकलेट  
[he bucket of dirty water from his chocolates  
निकालने ही वाला था] कि {उसकी मम्मी ने  
taking-out just doing was] that {his mother ERG  
उसे रोक दिया }  
him stop did}

“He was just going to take out the chocolates from the dirty water in the bucket when his mother stopped him.”

The method of collecting connectives will therefore necessarily involve “discovery during annotation”. However, we wanted to get some initial ideas about what kinds of connectives were likely to occur in real text, and to this end, we looked at 9 short stories with approximately 8000 words. Our goal here is to develop an initial set of guidelines for annotation, which will be done on the same corpus on which the sentence-level dependency annotation is being carried out (see Section 4). Table 1 provides the full set of connectives we found in our texts, grouped by syntactic type. The first four columns give the syntactic grouping, the Hindi connective expressions, the English gloss, and the English equivalent expressions, respectively. The last column gives the number of occurrences we found of each expression. In the rest of this section, we describe the function and distribution of discourse connectives in Hindi based on our texts. In the discussion, we have noted our points of departure from the PDTB where applicable, both with respect to the types of relations being annotated as well as with respect to terminology. For argument naming, we use the PDTB convention: the clause with which the connective is syntactically associated is called Arg2 and the other clause is called Arg1. Two special conventions are followed for paired connectives, which we describe below. In all Hindi examples in this paper, Arg1 is enclosed in square brackets and Arg2 is in braces.

<sup>4</sup> S=Subject; IO=Indirect Object; DO=Direct Object; V=Verb; ERG=Ergative; DAT=Dative

Connective Type	Hindi	Gloss	English	Num
Sub. Conj.	क्योंकि	why-that	because	2
	(क्यों)कि..इसलिए	(why)-that..this-for	because	3
	(अगर यदी)..तब तो	(if)..then	if..(then)	15
	(जब).. तब तो	(when)..then	when	50
	जब तक.. तब तक (के लिए)	when till..then till (of for)	until	2
	जैसे ही..(तो)	as just..(then)	as soon as	5
	इतना ऐसा..की	so such..that	so that	12
	ताकि कि	so-that that	so that when	1 5
Sentential Relatives	जिससे	which-with	because of which	5
	जो	which	because of which	1
	जिसके कारण	which-of reason	because of which	1
Subordinator	पर	upon	upon	9
	(-कर -के करके)	(do)	after while	111
	समय	time	while	1
	हुए	happening	while	28
	के बाद	of later	after	3
	से	with	due to	1
	के पहले	of before	before	1
	के लिए	of for	in order to	4
	में	in	while	1
के कारण	of reason	because of	3	
Coord. Conj.	लेकिन पर परन्तु	but	but	51
	और तथा	and	and	117
	या	or	or	2
	यों तो..पर	such TOP..but	but	2
	ना केवल..बल्कि	not only..but	not only..but	1
Adverbial	तब	then	then	2
	बाद में	later in	later	5
	फिर	then	then	4
	इसीलिए	this-for	that is why	7
	नहीं तो	not then	otherwise	5
	तभी तो	then-only TOP	that is why	1
	सो	so	so	10
	वही यही नहीं	that this-only not	not only that	1
<b>TOTAL</b>				<b>472</b>

Table 1: A Partial List of Discourse Connectives in Hindi. Parentheses are used for optional elements; “|” is used for alternating elements; TOP = topic marker.

### 3.1 Types of Discourse Connectives

#### 3.1.1 Subordinating Conjunctions

Finite adverbial subordinate clauses are introduced by independent lexical items called “subordinating conjunctions”, such as *क्योंकि* (“because”), as in (4), and they typically occur as right or left attached to the main clause.

- (4) मैं इस सभी धन को राज्य के बादशाह  
[I this all wealth ACC kingdom of king  
को दे देता], *क्योंकि* {वही समस्त

DAT give would], *why-that* {he-EMPH all  
धरती की सम्पदा का स्वामी है}  
earth of wealth of lord is}

“I would give all this wealth to the king, because he alone is the lord of this whole world’s wealth.”

As the first group in Table 1 shows, subordinating conjunctions in Hindi often come paired, with one element in the main clause and the other in the subordinate clause (Ex.5). One of these elements can also be implicit (Ex.6),

and in our texts, this was most often the subordinate clause element.

- (5) क्योंकि {यह तुम्हारी ज़मीन पर मिला है}, इसलिए  
because {this your land on found has}, this-for  
[इस धन पर तुम्हारा अधिकार है]  
[this treasure on your right is]

“Because this was found on your land, you have the right to this treasure.”

- (6) [उसका वश चलता] तो {वह उसे घर से  
[her power walk] then {she it home from  
बाहर निकाल देती}  
out take would}

“Had it been in her power, she would have banished it from the house.”

When both elements of the paired connective are explicit, their text spans must be selected discontinuously. The main clause argument is called Arg1 and the subordinate clause argument, Arg2.

Subordinating conjunctions, whether single or paired, can occur in non-initial positions in their clause. However, this word order variability is not completely unconstrained. First, not all conjunctions display this freedom. For example, while ‘जब’ (‘when’) can be clause-medial (Ex. 7), ‘क्योंकि’ (‘because’) cannot. Second, when the main clause precedes the subordinate clause, the main clause element, if explicit, cannot appear clause-initially at all. Consider the causal ‘क्योंकि.. इसलिये’ (Ex.5), which represents the subordinate-main clause order. In the reverse order, the explicit main clause ‘इसलिये’ (Ex.8) appears clause medially. Placing this element in clause-initial position is not possible.

- (7) {लकड़हारे की पत्नी को} जब {यह  
{woodcutter of wife DAT} when {this  
मालूम पड़ा कि इस चिड़िया के कारण  
knowledge put that this bird of reason  
काम छोड़कर घर आ गया है} तो [वह  
work leaving home come went is} then [she  
उस पर बरस पड़ी].  
him on anger-rain put}

“When the woodcutter’s wife found out that he had left his work and come home to care for the bird, she raged at him.”

- (8) [. . . पर चिराग की बत्ती उसका या दोहरी  
[. . .but lamp of light light or another  
बत्ती लगाना] शायद इसलिए [उचित नहीं

light putting] perhaps this-for [appropriate not  
समझते थे] कि {तेल का अपव्यय होगा}.

Consider did] that {oil of waste be-FUT}.

“... but he did not consider it appropriate to light the lamp repeatedly or light another lamp, perhaps because it would be a waste of oil.”

### 3.1.2 Sentential Relative Pronouns

Since discourse relations are defined as holding between eventualities, we have also identified relations that are expressed syntactically as relative pronouns in sentential relative clauses, which modify the main clause verb denoting an eventuality, rather than some entity denoting noun phrase. For example, in (9), a result/purpose relation is conveyed between ‘the man’s rushing home’ and ‘the bird being taken care of’, and we believe that this relation between the eventualities should be captured despite its syntactic realization as the relative pronoun ‘जिससे’ (‘because of which/so that’). (10) gives an example of a modified relative pronoun.

- (9) [सारा काम छोड़कर वह उस बीमार चिड़िया  
[all work leaving he that sick bird  
को उठाकर दवा घर की ओर भागा],  
ACC picking-up fast home of direction ran],  
जिससे {उसका सही इलाज किया जा सके}  
from-which {her proper care do go able}

“Leaving all his work, he picked up the bird and ran home very fast, so that the bird could be given proper care.”

- (10) [ऊँटों के हर वार कदम रखने पर  
[camels of every time step keeping upon  
चिड़ियों के सिर आपस में तथा ऊँट की  
birds of head each-other in and camels of  
गरदन से टकरा रहे थे] जिसके कारण  
neck with hit-against be had] of-which reason  
{उन पक्षियों की दरदभरी चीखें निकल  
{those birds of painful screams come-out  
रही थीं}.  
be had}

“With each step of the camels, the birds heads were hitting against each other as well as with the camels’ necks because of which the birds were screaming painfully.”

### 3.1.3 Subordinators

In contrast to the subordinating conjunctions, elements introducing non-finite subordinate clauses are called “subordinators”. Unlike

English, where certain non-finite subordinate clauses, called “free adjuncts”, appear without any overt marking so that their relationship with the main clause is unspecified, Hindi non-finite subordinate clauses almost always appear with overt marking. However, also unlike English, where the same elements may introduce both finite and non-finite clauses (cf. *After leaving, she caught the bus* vs. *After she left, she caught the bus*), different sets of elements are used in Hindi. In fact, as can be seen in the subordinator group in Table 1, the non-finite clause markers are either postpositions (Ex.11), particles following verbal participles (Ex.12), or suffixes marking serial verbs (Ex.13).

- (11) {मम्मी के मना करने} के कारण [रामू  
 {mummy of warning doing} of reason [Ramu  
 थोड़ी थोड़ी चॉकलेट बड़े अनंद के साथ  
 little little chocolate big pleasure of with  
 खा रहा था].  
 eat being be]

“Because of his mother’s warning, Ramu was eating bits of chocolate with a lot of pleasure.”

- (12) . . . और {खेलते} हुए [यह भूल जाता है  
 . . . and {playing} happening [this forget go is  
 कि यदि उसका मित्र भी अपने खिलौने को  
 that if his friends also their toys to  
 उसे हाथ नहीं लगाने देता, तो उसे  
 him hand not touching did, then he  
 कितना बुरा लगता]  
 how-much bad feel]

“. . . and while playing, he forgets that if his friends too didn’t let him touch their toys, then how bad he would feel.”

- (13) {अपनी पत्नी से यह सुन}कर [लकड़हारा  
 {self wife from this listen}-do [woodcutter  
 बहुत दुखी हुआ]  
 much sad became]

“Upon hearing this from his wife, the woodcutter became very sad.”

While subordinators constitute a frequently-used way to mark discourse relations, their annotation raises at least two difficult problems, both of which have implications for the reliability of annotation. The first is that these markers are used for marking both argument clauses and adjunct clauses, so that annotators would be required to make difficult decisions for distinguishing them: in the former case, the

marker would not be regarded as a connective, while in the latter case, it would. Second, the clauses marked by these connectives often seem to be semantically weak. This is especially true of verbal participles, which are nonfinite verb appearing in a modifying relation with another finite verb. Whereas in some cases (Ex.12-13) the two verbs are perceived as each projecting “two distinct events” between which some discourse relation can be said to exist, in other cases (Ex.14), the two verbs seem to project two distinct actions but as part of a “single complex event” (Verma, 1993). These judgments can be very subtle, however, and our final decision on whether to annotate such constructions will be made after some initial annotation and evaluation.

- (14) {देखते ही देखते सब बैल भागते }  
 {looking EMPH looking all buffalos running}  
 हुए [गोशाला पहुँच गए]  
 happening [shed reach did]

“Within seconds all the buffalos came running to the shed.”

The naming convention for the arguments of subordinators is the same as for the subordinating conjunctions: the clause associated with the subordinator is called Arg2 while its matrix clause is called Arg1.

Unlike subordinating conjunctions, subordinators do not come paired and they can only appear clause-finally. Clause order, while not fixed, is restricted in that the nonfinite subordinate clause can appear either before the main clause or embedded in it, but never after the main clause.

### 3.1.4 Coordinating Conjunctions

Coordinating conjunctions in Hindi are found in both inter-sentential (Ex.15) and intra-sentential (Ex.16) contexts, they always appear as independent elements, and they almost always appear clause-initially.<sup>5</sup> For these connectives,

<sup>5</sup> While the contrastive connectives ‘पर’, ‘परन्तू’ appear only clause-initially, it seems possible for the contrastive ‘लेकिन’ to appear clause-medially, suggesting that these two types may correspond to the English ‘but’ and ‘however’, respectively. However, we did not find any examples of clause-medial ‘लेकिन’ in our texts, and this behavior will have to be verified with further annotation.

the first clause is called Arg1 and the second, Arg2.

- (15) [जब वह लौटता तो गा-गाकर उसका मन  
[when he return then sing-singing his mind  
खुश कर देती]. लेकिन {उसकी पत्नी को वह  
happy do gave}. But {his wife DAT the  
चिड़िया फूटी आँख नहीं सुहाती थी}.  
bird torn eye not bear did}

“Upon his return, she would make him happy by singing. But his wife could not tolerate the bird even a little bit.”

- (16) [ तभी दरवाज़ा खुला] और {मालकिन आ  
[then-only door opened] and {wife come  
गई }.  
went}

“Just then the door opened and the wife came in.”

We also recognize paired coordinating conjunctions, such as ‘ना केवल..बल्कि’ (See Table 1). The argument naming convention for these is the same as for the single conjunctions.

### 3.1.5 Discourse Adverbials

Discourse adverbials in Hindi modify their clauses as independent elements, and some of these are free to appear in non-initial positions in the clause. Example (17) gives an example of the consequence adverb, ‘सो’. The Arg2 of discourse adverbials is the clause they modify, whereas Arg1 is the other argument.

- (17) [चिड़िया जबान कट जाने और मालकिन के ऐसे  
[bird tongue cut going and wife of this  
व्यवहार से डर गई थी]. सो {वह किसी  
behavior with fear go had}. So {she some  
तरह उड़कर चली गई}.  
manner flying walk went}.

“The bird was scared due to her tongue being cut and because of the wife’s behavior. So she somehow flew away.”

As with the PDTB, one of our goals with the Hindi discourse annotation is to explore the structural distance of Arg1 from the discourse adverbial. If the Arg1 clause is found to be non-adjacent to the connective and the Arg2 clause, it may suggest that adverbials in Hindi behave anaphorically. In the texts we looked at, we did not find any instances of non-adjacent Arg1s.

Additional annotation will provide further evidence in this regard.

## 4 Hindi Sentence-level Annotation and Discourse Connectives

The sentence-level annotation task in Hindi is an ongoing effort which aims to come up with a dependency annotated treebank for the NLP/CL community working on Indian languages. Presently a million word Hindi corpus is being manually annotated (Begum et al., 2008). The dependency annotation is being done on top of the corpus which has already been marked for POS tag and chunk information. The scheme has 28 tags which capture various dependency relations. These relations are largely inspired by the Paninian grammatical framework. Given below are some relations, reflecting the argument structure of the verb.

- a) कर्ता (agent) (k1)
- b) कर्म (theme) (k2)
- c) करण (instrument) (k3)
- d) सम्प्रदान sampradaan (recipient) (k4)
- e) अपादान (source) (k5)
- f) अधिकरण (location) (k7)

Figure 1 shows how Examples (2a-f) are represented in the framework. Note that agent and theme are rough translations for ‘कर्ता’ and ‘कर्म’ respectively. Unlike thematic roles, these relations are not purely semantic, and are motivated not only through verbal semantics but also through vibhaktis (postpositions) and TAM (Tense, aspect and modality) markers (Bharati et al., 1995). The relations are therefore syntactico-semantic, and unlike thematic roles there is a greater binding between these relations and the syntactic cues.

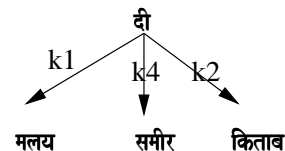


Figure 1: Dependency Diagram for Example (2) Some discourse relations that we have identified are already clearly represented in the sentence-level annotation. But for those that aren’t, the

discourse level annotations will enrich the sentence-level. In the rest of this section, we discuss the representation of the different types of connectives at the sentence level, and discuss how the discourse annotation will add to the information present in the dependency structures.

**Subordinating Conjunctions** Subordinating conjunctions are lexically represented in the dependency tree, taking the subordinating clause as their dependents while themselves attaching to the main verb (the root of the tree). Figure 2 shows the dependency tree for Example (4) containing the subordinating conjunction 'क्योंकि'. Note that the edge between the connective and the main verb gives us the causal relation between the two clauses, the relation label being 'rh' (relation hetu 'cause'). Thus, the discourse level can be taken to be completely represented at the sentence-level.

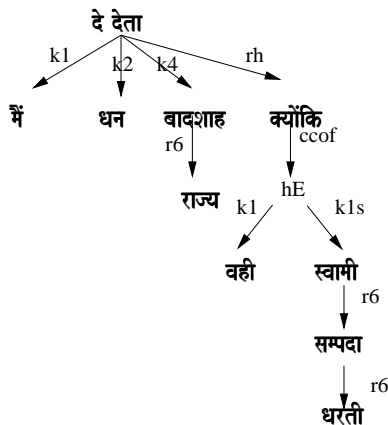


Figure 2: Dependency Tree for Subordinating Conjunction in Example (4)

**Paired Subordinating Conjunctions** Unlike Example (4), however, the analysis for the paired connective in Example (5), given in Figure 3, is insufficient. Despite the lexical representation of the connective in the tree, the correct interpretation of the paired conjunction and the clauses which it relates is only possible at the discourse level. In particular, the dependencies don't show that 'क्योंकि' and 'इसलिए' are two parts of the same connective, expressing a single relation and taking the same two arguments. Thus, the discourse annotation will

be able to provide the appropriate argument structure and semantics for these paired connectives.

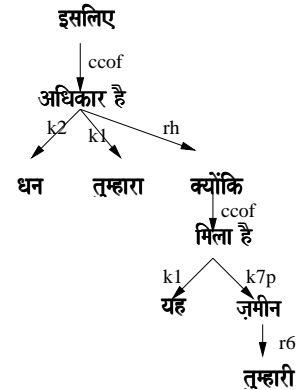


Figure 3: Dependency Tree for Paired Subordinating Conjunction in Example (5)

**Subordinators** As mentioned earlier, Hindi nonfinite subordinate clauses almost always appear with overt marking. But unlike the subordinating conjunctions, subordinators are not lexically represented in the dependency trees. Figure 4 gives the dependency representation for Example (11) containing a postposition subordinator 'के कारण', which relates the main and subordinate clauses causally. As the figure shows, while the causal relation label ('rh') appears on the edge between the main and subordinate verbs, the subordinator itself is not lexically represented as the mediator of this relation. The lexically grounded annotation at the discourse level will thus provide the textual anchors of such relations, enriching the dependency representation. Furthermore, while many of the subordinators in Table 1 are fully specified in the dependency trees for the semantic relation they denote (e.g., 'पर' and 'में' marked as the 'k7t' (location in time) relation, and 'के कारण' and 'से' marked as the 'rh' (cause/reason) relation), others, like the particle 'हुए' are underspecified for their semantics, being marked only as 'vmod' (verbal modifier). The discourse-level annotation will thus be the source for the semantics of these subordinators.

**Coordinating Conjunctions** Coordinating conjunctions at the sentence level anchor the root of the dependency tree. Figure 5 shows the

dependency representation of Example (16) containing a coordinating conjunction.

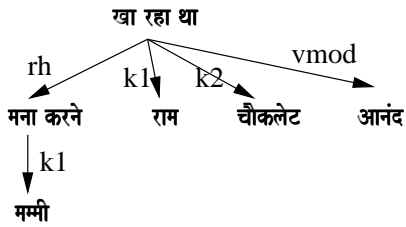


Figure 4: Dependency Tree for Subordinator in Example (11)

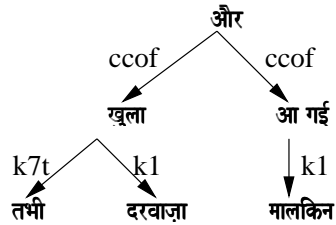


Figure 5: Dependency Tree for Coordinating Conjunction in Example (16)

While the sentence-level dependency analysis here is similar to the one we get at the discourse level, the semantics of these conjunctions are again underspecified, being all marked as ‘ccof’, and can be obtained from the discourse level.

**Discourse Adverbials** Like subordinating conjunctions, discourse adverbials are represented lexically in the dependency tree. They are attached to the verb of their clause as its child node and their denoted semantic relation is specified clearly. This can be seen with the temporal adverb ‘तभी’ (‘then-only’) and its semantic label ‘k7t’ in Figure 5. At the same time, since the Arg1 discourse argument of adverbials is most often in the prior context, the discourse annotation will enrich the semantics of these connectives by providing the Arg1 argument.

## 5 Summary and Future Work

In this paper, we have described our study of discourse connectives in a small corpus of Hindi texts in an effort towards developing an annotated corpus of discourse relations in Hindi. Adopting the lexically grounded approach of the Penn Discourse Treebank, we have identified a

wide range of connectives, analyzing their types and distributions, and discussing some of the issues involved in the annotation. We also described the representation of the connectives in the sentence-level dependency annotation being carried out independently for Hindi, and discussed how the discourse annotations can enrich the information provided at the sentence level. While we focused on explicit connectives in this paper, future work will investigate the annotation of implicit connectives, the semantic classification of connectives, and the attribution of connectives and their arguments.

## References

- Rafiya Begum, Samar Husain, Arun Dhawaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of IJCNLP-2008*. Hyderabad, India.
- Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice Hall of India. <http://ltrc.iiit.ac.in/downloads/nlpbook/nlppanini.pdf>.
- Manindra K. Verma (ed.). 1993. *Complex Predicates in South Asian Languages*. New Delhi: Manohar.
- The PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS-06-01, IRCS, University of Pennsylvania.
- Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.
- Nianwen Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, Michigan.
- Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In *Proceedings of IJCNLP-2008*. Hyderabad, India.



# A Semantic Study on Yami Ontology in Traditional Songs

**Yin-Sheng Tai**  
Providence University,

Taiwan

[wratp2@msn.com](mailto:wratp2@msn.com)

**D. Victoria Rau**  
Providence University,

Taiwan

[dhrau@pu.edu.tw](mailto:dhrau@pu.edu.tw)

**Meng-Chien Yang**  
Providence University,

Taiwan

[mcyang2@pu.edu.tw](mailto:mcyang2@pu.edu.tw)

## Abstract

The purpose of this study was to provide an example of how to build a Yami ontology from traditional songs by employing Protégé, an open-source tool for editing and managing ontologies developed by Stanford University. Following Conceptual Blending Theory (Fauconnier and Turner, 1998), we found that Yami people use the conceptual metaphor of “fishing” in traditional songs when praising the host’s diligence in a ceremony celebrating the completion of a workhouse. The process of building ontologies is explored and illustrated. The proposed construction of an ontology for Yami traditional songs can serve as a fundamental template, using the corpus available online from the Yami documentation website (<http://yamiproject.cs.pu.edu.tw/yami>) to build ontologies for other domains.

## 1 Introduction

Yami is an endangered Austronesian language, spoken on Orchid Island (Lanyu), 46 kilometers southeast of the main island of Taiwan. For the purposes of language documentation and preservation, an on-line Yami dictionary<sup>1</sup> has been developed to facilitate language learning. Although each lexical entry contains basic meanings of words (in both English and Chinese), pronunciations, and roots and affixes, no information on lexical semantics, such as synonyms, hyponyms, or metaphors is available. If information on lexical relationships could be incorporated into the Yami dictionary, this online tool would be even more useful for Yami language learners.

In the present study, we focused on the metaphors in Yami lyrics. Knight (2005) considers that, from a Yami native speaker’s point of view,

<sup>1</sup> Available from the following IP address:  
<http://yamiproject.cs.pu.edu.tw/elearn/search.php>

“Raods” (traditional songs) play an important role in culture because they subsume features such as archaism, metaphors, puns and polite contradiction. In addition, Yami traditional songs reflect Yami values, such as love and honoring hard work (e.g. fishing or farm work), and cultural events, such as completion of hard work and special festivals. Thus, we would like to build an ontology using Yami traditional songs, adapting the taxonomies in WordNet and SUMO, which can serve as a point of departure for further mapping of other Yami ontologies. In this paper, we will report our preliminary results, giving one example at this early stage of the research project on constructing Yami ontology, led by the second and third authors.

## 2 Literature Review

### 2.1 Conceptual Metaphor Theory

The original Conceptual Metaphor Theory was proposed by Lakoff and Johnson (1980). They identify metaphor as a transfer between the source domain and the target domain. This has become known as the “two-domain theory” of metaphor.

### 2.2 Conceptual Blending Theory

Conceptual Blending Theory (Fauconnier and Turner, 1998; 2002) is a framework for interpreting cognitive linguistic phenomena such as analogy, metaphor, etc. According to Conceptual Blending Theory, the input structures, generic structures, and blend structures in the network are *mental spaces*. In Figure 1, the frame structure recruited to the mental space is represented as a rectangle either outside or iconically inside the circle.

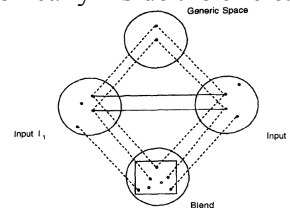


Figure 1. Conceptual Blending Theory

### 2.3 Protégé

In this study, we are building a Yami ontology based on traditional songs using Protégé<sup>2</sup>, which not only provides a rich set of basic knowledge-modeling structures and a way to enter data, but can also be customized to create new domains in knowledge models.

As demonstrated in previous studies (e.g., Dodds, 2005; Lin 2006), Protégé has been successfully applied to construction of ontology in specific domains. Therefore, the present study used Protégé to construct an ontology of Yami traditional songs.

## 3. Methodology

### 3.1 Data Collection

The main data resource came from Dong's monograph (1994) *"In Praise of Taro,"* which contains a total of 250 songs. This study is based on one song which contains many fishing metaphors. Seven metaphorical tokens were extracted from this song.

### 3.2 Data Analysis

First of all, the question of the use of metaphor in Yami was analyzed by Lakoff & Johnson's Conceptual Metaphor Theory (1980) and Fauconnier and Turner's Conceptual Blending Theory (1998). Secondly, two taxonomic tenors were identified using *WordNet* (Fellbaum, 1999) and *Yami Texts with Reference Grammar and Dictionary* (Rau & Dong, 2006). Based on these taxonomic tenors, Yami words were classified into "Verbs" and "Nouns".

## 4. Results and Discussion

### 4.1 Conceptual Blending

In the following discussion, we begin with an analysis of a traditional Yami song celebrating the completion of a workhouse with a harvest of taro. It mostly praises the host's achievement and hard work. After praising the host's achievement, the guests take all the host's taros and cover the roof of the workhouse with them. Finally, the guests sing songs with the host in turn.

The lyric is illustrated as follows:

1 oya rana **minangyid** siapen rarakeh  
this already reached\_the\_harbor grandfather old

*"Now, the old man reached the harbor."*

2 ji na minatokod **Jicamongan** ta  
NEG already reached PLN because  
*"He didn't paddle to Jicamongan."*

3 **kalagarawan** am **paneneneban** o ...  
fingerling\_place TOP shallow\_sea NOM ...  
*"He moved in the shallow sea where only fingerling fish live."*

4 to na rana **avavangi** sia ta  
AUX 3.S.GEN already row a boat there because  
*"He could only row his boat there"*

5 ji na rana **voaz** o **kakaod**  
NEG 3.S.GEN already row NOM paddle  
*"Because he had already lost strength to row."*

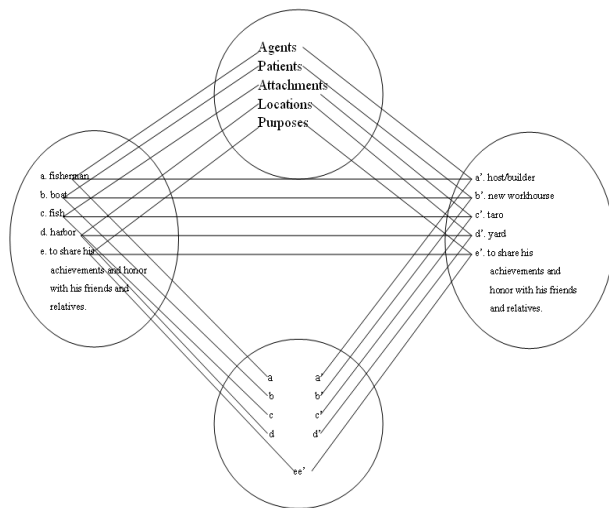
Seven metaphorical tokens related to "fishing" were detected from the lyric. They are marked in bold case. In Line 1, *minangyid* "reached the harbor" was identified as metaphorical because although it literally describes going back home after finishing one's fishing, its intended meaning is "to rest and hold a ceremony celebrating the completion of a workhouse." Using the Conceptual Metaphor Theory (1980), we compared the cognitive activities in the lyrics and the mental space (Table 1). The entities, quality, and functions in the domain of fishing were analyzed.

**Table 1. Reached the harbor vs. Holding a ceremony celebrating the completion of a workhouse**

	reached the harbor (V.)	to hold a ceremony celebrating the completion of a workhouse
Agents	fisherman	host (also the "builder") of the ceremony
Patients	boat	new workhouse
Attachments	fish	taro
Locations	harbor	yard
Purposes	to share his achievements and honor with his friends and relatives	to share his achievements and honor with his friends and relatives

Table 1 shows that Yami people prefer to use "fishing" as a metaphor for the intended meanings of building a house or farming. We further employed the Conceptual Blending Theory (1998) to interpret the "harbor" example (see Figure 2).

<sup>2</sup> Protégé is a free, open source ontology editor and knowledge-base framework. The IP address of Protégé is: <http://protege.stanford.edu/>.



**Figure 2. “Reached the harbor” vs. “Holding a ceremony celebrating the completion of a workhouse”**

The concept of “reached the harbor” is categorized into Input space I, and “holding a ceremony celebrating the completion of a workhouse” is categorized into Input space II. Both Inputs have certain similarities as well as distinct features. From Input Space I, “fisherman,” “boat,” “fish,” “harbor,” and “to share his achievements and honor with his friends and relatives” are respectively mapped into “host/builder,” “new workhouse,” “taro,” “yard,” and “to share his achievements and honor with his friends and relatives,” in Input Space II. The Inputs might share some cross-mapping properties, which can be listed in the Generic space. The structure from the two input mental spaces is projected into the Blend space. Essentially, which elements from Input space II should be selected and projected onto the blend space are determined by the contents of the lyric. Thus, in the blend space, all elements remain separate from their corresponding counterparts, but the relations among the features in Input space I determines the relations between corresponding counterparts. That is to say, the running structure in the blend space partially projected from Input space I determines the existing relations among the elements in the blend. In Input space I, a “fisherman” needs a “boat” to fish in the ocean, so the relation between “fisherman” and “boat” is a kind of “earning a living.” In addition, what a “fisherman” works for is “fish.” After the fisherman finishes his work, he has to go back to the harbor. Such relations also operate among those

elements projected from Input space II. As a result, “the host of the completion ceremony of a workhouse” is the “fisherman” of the “workhouse ceremony;” the relation between “the host” and the “workhouse ceremony” is that of “finishing a time-consuming job.” Additionally, “taros” in the yard are compared with “fish” at the harbor, which await to be “shared with friends and relatives.”

#### 4.2 Taxonomy

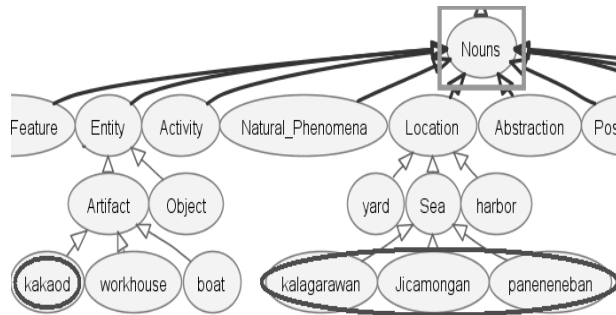
Yami verbs subsume dynamic verbs and stative verbs (Rau and Dong, 2006). Based on the notions from *WordNet*, we further divided verbs into Bodily Function and Care Verbs, Change Verbs, Communication Verbs, Competition Verbs, Consumption Verbs, Contact Verbs, Cognition Verbs, Creation Verbs, Motion Verbs, Emotion or Psych Verbs, Stative Verbs, Perception Verbs, Possession Verbs, Social Interaction Verbs, and Weather Verbs. Since Yami does not possess a distinctive adjective word class, both descriptive and relational adjectives are in this study classified under stative verbs in *WordNet*. Descriptive adjectives subsume antonymy, gradation, markedness, polysemy and selectional preferences, reference-modifying adjectives, color adjectives, quantifiers, and participial adjectives. Relational adjectives include two domains, “pertaining” or “relating to” (Fellbaum 1999: 63). The coding of adjectives in this file is different from that of descriptive adjectives. Rather than being part of a cluster, each synset is entered individually, so that the interface will present the adjective with its related noun and information about the sense of the noun.

For the aspect of Yami Nouns, we categorized the nouns into 10 basic noun categories following *WordNet* (Fellbaum, 1999: 30), including entity, abstraction, psycho-feature, natural phenomena, activity, event, group, location, possession, and state.

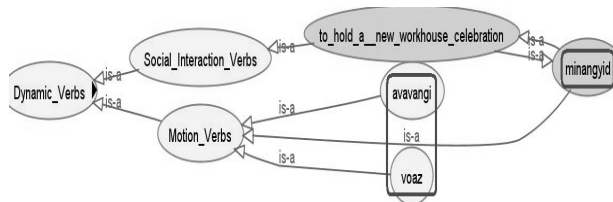
#### 4.3 Example from the Yami Lyric

The following example illustrates how we extracted the metaphorical words of “fishing” from the lyric and classified them into their domains. Firstly, *minangvid* “reached the harbor,” *avavangi* “row, sail something” and *voaz* “row something,” were classified as Motion Verbs. Secondly, *Jicamongan* “a place name of deep water,” *kalagarawan* “a place where fingerling fish swim” and *paneneneban* “shallow place,” were categorized under the main section Location and

the sub-section Sea. Finally, *kakaod* “paddle” was classified in the main section Entity and the sub-section Artifact, as shown in Figure 3 and Figure 4.



**Figure 3. An example of four Yami nouns in the ontology of Yami lyrics**



**Figure 4. An example of three Yami verbs in the ontology of Yami lyrics**

#### 4.4 Mapping metaphorical words

To classify Yami metaphor, we made links (equivalent classes) between the literal meaning and the metaphorical meaning using Protégé. Since Input space I usually works as the source domain and Input space II as the target domain, we found that, at least in this case, this is a one-sided network of metaphor mapping. We thus employed the property “mappings” to define the relationship between “reached the harbor” and “to hold a new workhouse celebration.” Moreover, in order to set up the restrictions for searching words, we added the property “hasConceptOf”. We then named the class expression of “reached the harbor” with the following restriction:

*has\_concept\_of\_some<sup>3</sup> (harbor and<sup>4</sup> fish and fisherman and boat and share)*

Similarly, we also provided the phrase “to hold a new workhouse celebration” with the following restriction:

*has\_concept\_of\_some (host and builder and workhouse and share and taro and yard)*

Finally, for the sake of correctly linking

<sup>3</sup> This refers to the existential quantifier ( $\exists$ ) in OWL syntax, which can be read as at least one, or *some*.

<sup>4</sup> An intersection class is described by combining two or more classes using the *AND* operator ( $\sqcap$ ).

meanings of each word, the ontology builder can check the ontology with the DL Query Tab in Protégé.

In summary, the connection between the two items *minangyid*, “reached the harbor” and “to hold a new workhouse celebration,” shown in shadow in Figure 4, is solely based on the structure of the inputs, since each of them are from a different domain of verbs. This structure is in harmony with Yami custom and contextual structure.

## 5 Conclusion

This paper has provided an example of how to construct an ontology of Yami using traditional songs. We hope this approach will serve as a fundamental template for further mappings with more texts to produce other Yami ontologies.

## References

- Dodds, D. 2005. *Qualitative geospatial processing, ontology and spatial metaphor*. presented at the GML and Geo-Spatial Web Services.
- Dong, M. N. 1995. *In Praise of Taro*. Dao-Xiang Publisher.
- Fauconnier, G. & Turner, M. 1998. Conceptual Integration Networks. *Cognitive Science*, 22 (2), 133-187.
- Fellbaum, C. 1999. *WordNet: an Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Knight, P., & Lu Y. H.. 2005. *Music heritage of the oral traditions by meykaryag of the Tao tribe*. Paper presented at the 2005 International Forum of Ethnomusicology in Taiwan: Interpretation and Evolution of Musical Sound. Taipei: Soochow University.
- Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.
- Lin, J.H. 2006. *Using formal concept analysis to construct the computer virus characteristics domain ontology*. MA thesis, National Yunlin University of Science & Technology, Taiwan.
- Rau, D. Victoria & Dong, M. N. 2006 *Yami Texts with Reference Grammar and Vocabulary*. Language and Linguistics, Academia Sinica, Taipei, Monograph A-10.

# ASSESSMENT AND DEVELOPMENT OF POS TAG SET FOR TELUGU

Dr.Rama Sree R.J  
Rashtriya Sanskrit  
Vidyapeetha,  
Tirupati  
[rjramasree@yahoo.com](mailto:rjramasree@yahoo.com)

Dr.Uma Maheswara Rao G  
Central University  
Hyderabad  
[guraohyd@yahoo.com](mailto:guraohyd@yahoo.com)

Dr. Madhu Murthy K.V  
S.V.U.College of  
Engineering  
Tirupati  
[kvmmurthy@yahoo.com](mailto:kvmmurthy@yahoo.com)

## ABSTRACT

In this paper, we first had a overall study of existing POS tag sets for European and Indian languages. Till now, most of the research done on POS tagging is for English. We observed that even though the research on POS tagging for English is done exhaustively, part-of-speech annotation in various research applications is incomparable which is variously due to the variations in tag set definitions. We understand that the morpho-syntactic features of the language and the degree of desire to represent the granularity of these morpho-syntactic features, domain etc., decide the tags in the tag set. We then examined how POS tagset design has to be handled for Indian languages, taking Telugu language into consideration.

## 1. Introduction

Annotation is the process of adding some additional information (grammatical features like word category, case indicator, other morph features) about the word to each word of the text. This additional information is called a tag. The set of all these tags is called a tag set. When words are considered in isolation, they can have one or more number of tags for each word. But when these words are used in a certain context, the tags representing morphological and syntactic feature reduce to one tag. The information to be captured as a tag is an application specific issue (Anne,1997, David, 1994 and David,

1995). A number of tag sets have been evolved for a number of languages. These tag sets not only differ with each other from language to language, but vary within the language itself. The reasons for the variation of tags in the tag sets are as follows. As taggers give additional information like grammatical features such as number, gender, person, case markers for noun inflections; tense markers for verbal inflections, the number of tags used by different systems varies depending on the information encoded in the tag. However the tag set design plays a vital role when data is tagged according to it and hence it affects the development of NLP application tools within and across that language. Language independent representation of a tag set help to find out the hidden information like context, structure, syntactic and semantic aspect of the word. It also gives an overview of language modeling features.

## 2. Desirable Features of a Tag Set

Unfortunately, there does not seem to be much literature on standard tag set design. There is a need to have standard tag set labels for the words to encode the same linguistic information across the languages. The tag set labels of a given language should satisfy the following characteristics.

- (1) The words carrying same syntactic, categorical information should be grouped under the same tag. For example, all adjectives should be tagged as JJ.

- (2) The words which have same syntax and come under different categories should clearly be distinguished depending on the categorical sense in which it is used in the given context. For example, the word **book** can be tagged both as noun (NN) and Verb (VB).
- (3) The tag set should also help us to classify and predict the sense, category of the unknown and foreign words. For example, consider the sentence, “Give it to **xyxy**”, POS tagger should be in a position to predict **xyxy** (or any non-sensical string) could be a noun.

### 3. Sources of Variations among POS Tag Sets for English

In order to identify the reasons for tag set variations for English, the tag sets viz., the Penn Treebank (Mitchel, 1993) tag set (PT), UCREL CLAWS7 tag set (UCREL\_C7), the International Corpus of English (ICE) tag set (Greenbaum, 1992) and the Brown Corpus (BC) tag set (Green, 1997) for English are examined; the POS tag labels are extracted for some important morpho-syntactic features and studied to demonstrate the present study.

After a careful study, the following points were observed with regard to the differences in POS tag sets.

- (i) **Desire to capture more semantic content:** BC, ICE, URCEL tag set are making more subtle distinctions within one category than PT. For example, POS tags for adjectives- PT is not making any clear distinction for adjectives other than JJ, JJS, JJR, whereas other tag sets are maintaining fine granularity. Such differences can be observed for several morpho-syntactic features.
- (ii) **Corpus Coverage:** Depending on the syntactic distribution of the test corpus under consideration, there may be variations. For example, BC tag set made a wide provision for foreign words (not

shown in the above table). In British corpus, there may be a possibility of the presence of the test corpus where more number of words are borrowed from other languages into English.

- (iii) **Desire for precision:** The reason for more number of tags in a tag set is to precisely capture all linguistic criteria which describe morpho-syntactic features in detail. However, there should be a balance between theoretical and actual distribution of these syntactic features.

### 4. Tag Sets for Indian Languages

The two POS tag sets developed for Hindi (revised on Nov 15, 2003) and Telugu by IIT, Hyderabad and CALTS, Hyderabad respectively are examined and the following points are observed.

Telugu POS tag set contains more number of POS tag labels. This difference is due to the reason that Telugu is more inflective than Hindi. In Hindi nouns are non-inflectional. *Karaka* roles are not encoded in Hindi noun word forms as in Telugu. Similarly main verbal roots appear as non-inflective in Hindi. The verbs co-occur with tense, aspect and modality as separate words whereas aspect and modality are packed into a single verbal inflection word in Telugu. For example, consider the following sentences.

English: **Ram killed Ravana.**

Hindi : **RAm ne mArA Ravana ko.**

Telugu: **RAmudu caMpAdu RAVanunni.**

For convenience, the word order is maintained as it is in all the three languages. In case of English language, position gives the roles played by Rama (subject) and Ravana (object). In case of Hindi, case markers *ne* and *ko* exist, but they do not inflect Ram and Ravana. But Telugu noun inflections give the information of case markers also. Hence there are differences in the tag labels of Hindi and

Telugu language tag sets. In order to capture these syntactic (more over they are also semantic) information, Telugu has more number of POS tags (nn1,nn2,nn3, nn4,nn5,nn6,nn7) in the place of a single tag (nn) of Hindi.

The POS tags of Telugu are described below in detail.

**(i) Nouns (nAma vAcakAlu- nn)** :These tags capture the nouns and their roles played in the sentence. The different tags in the subclass are *nn1,nn2,nn3,nn4,nn5, nn6* and *nn7*. Depending on the *vibhakti*, the nouns get the number label to main class, i.e., *nn* based on the *karaka* relations. The tag *nni* stands for noun oblique form indicating that the noun is in a position to get attached with the succeeding noun inflection.

**(ii) Locative affixes (swAna vAcakAlu – nl)** :Here some locative prepositions combined with the six *vibhaktis* are listed as *nl1* (pEna-పైన), *nl4* (pEki-పైకి), *nl5* (pEnuMdi-పైనుండి), *nl6* (pEni-పైని) etc.

**(iii) Prepositions (Vibhakti – pp)**:Sometimes prepositions can occur independently. For example, *varaku* (వరకు). Hence all *vibhaktis* are labelled as *pp1,pp2* etc.

**(iv) Pronouns (sarva nAmAlu - pr)** :Like nouns, all pronouns form inflections with *vibhaktis*. Accordingly they are named as *pr1, pr2* etc.

**(v) Adjectives (Visheshana)** : Special type of adjectives like Verbal adjectives ( *kriya visheshana*) as *vjj*, Nominal adjectives (*saMjna viseshana*) as *jj* and noninfinitive verbal adjectives (*sahAyaka asamapaka kriya*) as *ajj* etc.

**(vi) Other syntactic categories** : The tags for other syntactic categories like quantifiers as *qf*, negative meanings as *ng* etc., are given.

## 5. Improvement of Telugu Tag Set

In addition to the above mentioned tags, some new tags are introduced to capture and provide finer discrimination of the semantic content of some of the linguistic expressions a corpus of 12,000 words. They are explained briefly in the succeeding paragraphs.

**(a) Verbal finite negative** : Some words like *kAxu* (కాదు), *lexu* (లేదు) are verbal finites but they give the negative meaning of the verbal action. If they are tagged simply as *vf*, it is understood that some action has taken place. But these words are used in negative sense. In order to capture this feature, we have labelled them as **vng**.

**(b) Verbal nouns with vibhakti**: Verbal nouns behave in the same way as nouns do, in forming their inflections with *vibhaktis* like *Adatam*-(అడటం), *Adatanni*-(అడటాన్ని),

*Adatamcewa* (అడటం చేత) etc. At present they are labelled as *nn1, nn2, nn3* etc. depending on the affix. In doing so, the semantic content of verb is lost. This would lead to difficulties in disambiguating words at the semantic level. Hence the introduction of POS tags like *vnn1, vnn2* etc is proposed.

**(c) Words expressing doubts**: There are linguistic expressions that express the doubtfulness as explained below.

Doubtfulness of :

**(i) Verbal finites**:Words like *uMxo* (ఉందో) *vunnavo* (ఉన్నవో) etc., express the doubtfulness of the occurrence of action. To capture this semantic discrimination, POS tag *vw* is introduced. Previously they are labelled as *vf*.

**(ii) Nouns**: Words which express the doubtfulness a noun participation in the action like *rAmudo* (రాముడో), *axo* (అదో)

etc. Instead of labelling them *nnI*, they are labelled them with the tag *nnw*.

The above mentioned improvements made to the existing POS tag sets and the advantages thereof are as follows.

- (i) A finer discrimination is made. For example consider *vw*. In the absence of this tag, the verbal inflections which end with *lexu* (లేదు) could be tagged as *vf*.

Due to this, the verbal inflections which are completed can be clearly distinguished from those verbal inflections where action has not been completed.

- (ii) *vnn* tag captures more information that the noun present in the verbal inflection is just a simple common noun. In the absence of this tag, words erroneously labelled as *nn* to which it does not really belong. So these tags accurately capture the information present in the words.

## 6. Conclusion

It is strongly felt that all Indian languages should have the same tag set so that the annotated corpus in corresponding languages may be useful in cross lingual NLP applications, reducing much load on language to language transfer engines. This point can be well explained by taking analogy of existing script representation for Indian Languages. The ISCII and Unicode representations for all Indian languages can be viewed appropriately in the languages we like, just by setting their language code. There is no one-to-one alphabet mapping in the scripts of Indian Languages. For example, the short e,o (ఎ,ఐ)

are present in Telugu, while they are not available in Hindi, Sanskrit etc. Similarly alphabet variations between Telugu and Tamil exist. Even then, all these issues are taken care of, in the process of language to language script conversion. Similarly POS variations across Indian Languages also should be taken care of.

## References:

- Anne Schiller, Simone Teufel, Christine Thielen. 1995. **Guidelines für das Tagging deutscher Textcorpus mit STTS.** Universitäten Stuttgart und Tübingen.
- David Elworthy. 1994. **Automatic error detection in part of speech tagging.** In Proceedings of the International Conference on New Methods in Language Processing, Manchester.
- David Elworthy. 1995. **Tagset Design and Inflected Languages.** In Proceedings of the ACL SIGDAT Workshop, Dublin.
- Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz. 1993. **Building a Large Annotated Corpus of English: The Penn Treebank.** Computational Linguistics. Volume 19, Number 2, pp. 313--330 (Special Issue on Using Large Corpus).
- Greenbaum S. 1992. **The ICE tag set manual.** University College London.
- Green B, Rubun G. 1971. **Automated Grammatical Tagging of English.** In Department of Linguistics, Brown University.



## Designing a Common POS-Tagset Framework for Indian Languages

Sankaran Baskaran, Microsoft Research India, Bangalore. [baskaran@microsoft.com](mailto:baskaran@microsoft.com)

Kalika Bali, Microsoft Research India, Bangalore. [kalikab@microsoft.com](mailto:kalikab@microsoft.com)

Tanmoy Bhattacharya, Delhi University, Delhi. [tanmoy1@gmail.com](mailto:tanmoy1@gmail.com)

Pushpak Bhattacharyya, IIT-Bombay, Mumbai. [pb@cse.iitb.ac.in](mailto:pb@cse.iitb.ac.in)

Girish Nath Jha, Jawaharlal Nehru University, Delhi. [girishj@mail.jnu.ac.in](mailto:girishj@mail.jnu.ac.in)

Rajendran S, Tamil University, Thanjavur. [raj\\_ushush@yahoo.com](mailto:raj_ushush@yahoo.com)

Saravanan K, Microsoft Research India, Bangalore. [v-sarak@microsoft.com](mailto:v-sarak@microsoft.com)

Sobha L, AU-KBC Research Centre, Chennai. [sobha@au-kbc.org](mailto:sobha@au-kbc.org)

Subbarao K V, Delhi. [kvs2811@yahoo.com](mailto:kvs2811@yahoo.com)

### Abstract

Research in Parts-of-Speech (POS) tagset design for European and East Asian languages started with a mere listing of important morphosyntactic features in one language and has matured in later years towards hierarchical tagsets, decomposable tags, common framework for multiple languages (EAGLES) etc. Several tagsets have been developed in these languages along with large amount of annotated data for furthering research. Indian Languages (ILs) present a contrasting picture with very little research in tagset design issues. We present our work in designing a common POS-tagset framework for ILs, which is the result of in-depth analysis of eight languages from two major families, viz. Indo-Aryan and Dravidian. Our framework follows hierarchical tagset layout similar to the EAGLES guidelines, but with significant changes as needed for the ILs.

### 1 Introduction

A POS tagset design should take into consideration all possible morphosyntactic categories that can occur in a particular language or group of languages (Hardie, 2004). Some effort has been made in the past, including the EAGLES guidelines for morphosyntactic annotation (Leech and Wilson, 1996) to define guidelines for a common tagset across multiple languages with an aim to capture

more detailed morphosyntactic features of these languages.

However, most of the tagsets for ILs are language specific and cannot be used for tagging data in other language. This disparity in tagsets hinders interoperability and reusability of annotated corpora. This further affects NLP research in resource poor ILs where non-availability of data, especially tagged data, remains a critical issue for researchers. Moreover, these tagsets capture the morphosyntactic features only at a shallow level and miss out the richer information that is characteristic of these languages.

The work presented in this paper focuses on designing a common tagset framework for Indian languages using the EAGLES guidelines as a model. Though Indian languages belong to (mainly) four distinct families, the two largest being Indo-Aryan and Dravidian, as languages that have been in contact for a long period of time, they share significant similarities in morphology and syntax. This makes it desirable to design a common tagset framework that can exploit this similarity to facilitate the mapping of different tagsets to each other. This would not only allow corpora tagged with different tagsets for the same language to be reused but also achieve cross-linguistic compatibility between different language corpora. Most importantly, it will ensure that common categories of different languages are annotated in the same way.

In the next section we will discuss the importance of a common standard vis-à-vis the currently available tagsets for Indian languages. Section 3 will provide the details of the design principles

behind the framework presented in this paper. Examples of tag categories in the common framework will be presented in Section 4. Section 5 will discuss the current status of the paper and future steps envisaged.

## 2 Common Standard for POS Tagsets

Some of the earlier POS tagsets were designed for English (Greene and Rubin, 1981; Garside, 1987; Santorini, 1990) in the broader context of automatic parsing of English text. These tagsets popular even today, though designed for the same language differ significantly from each other making the corpora tagged by one incompatible with the other. Moreover, as these are highly language specific tagsets they cannot be reused for any other language without substantial changes this requires standardization of POS tagsets (Hardie 2004). Leech and Wilson (1999) put forth a strong argument for the need to standardize POS tagset for *reusability* of annotated corpora and *interoperability* across corpora in different languages. EAGLES guidelines (Leech and Wilson 1996) were a result of such an initiative to create standards that are common across languages that share morphosyntactic features.

Several POS tagsets have been designed by a number of research groups working on Indian Languages though very few are available publicly (IIT-tagset, Tamil tagset). However, as each of these tagsets have been motivated by specific research agenda, they differ considerably in terms of morphosyntactic categories and features, tag definitions, level of granularity, annotation guidelines *etc.* Moreover, some of the tagsets (Tamil tagset) are language specific and do not scale across other Indian languages. This has led to a situation where despite strong commonalities between the languages addressed resources cannot be shared due to incompatibility of tagsets. This is detrimental to the development of language technology for Indian languages which already suffer from a lack of adequate resources in terms of data and tools.

In this paper, we present a common framework for all Indian languages where an attempt is made to treat equivalent morphosyntactic phenomena consistently across all languages. The hierarchical design, discussed in detail in the next section, also allows for a systematic method to annotate lan-

guage particular categories without disregarding the shared traits of the Indian languages.

## 3 Design Principles

Whilst several large projects have been concerned with tagset development very few have touched upon the design principles behind them. Leech (1997), Cloeren (1999) and Hardie (2004) are some important examples presenting universal principles for tagset design.

In this section we restrict the discussion to the principles behind our tagset framework. Importantly, we diverge from some of the universal principles but broadly follow them in a consistent way.

**Tagset structure:** *Flat tagsets* just list down the categories applicable for a particular language without any provision for modularity or feature reusability. *Hierarchical tagsets* on the other hand are structured relative to one another and offer a well-defined mechanism for creating a common tagset framework for multiple languages while providing flexibility for customization according to the language and/ or application.

*Decomposability* in a tagset allows different features to be encoded in a tag by separate sub-stings. Decomposable tags help in better corpus analysis (Leech 1997) by allowing to search with an underspecified search string.

In our present framework, we have adopted the hierarchical layout as well as decomposable tags for designing the tagset. The framework will have three levels in the hierarchy with categories, types (subcategories) and features occupying the top, medium and the bottom layers.

**What to encode?** One thumb rule for the POS tagging is to consider only the aspects of morpho-syntax for annotation and not that of syntax, semantics or discourse. We follow this throughout and focus only on the morphosyntactic aspects of the ILs for encoding in the framework.

**Morphology and Granularity:** Indian languages have complex morphology with varying degree of richness. Some of the languages such as those of the Dravidian family also display agglutination as an important characteristic. This entails that morphological analysis is a desirable pre-process for the POS tagging to achieve better results in automatic tagging. We encode all possible morphosyntactic features in our framework assuming the exist-

tence of morphological analysers and leave the choice of granularity to users.

As pointed out by Leech (1997) some of the linguistically desirable distinctions may not be feasible computationally. Therefore, we ignore certain features that may not be computationally feasible at POS tagging level.

**Multi-words:** We treat the constituents of Multi-word expressions (MWEs) like *Indian Space Research Organization* as individual words and tag them separately rather than giving a single tag to the entire word sequence. This is done because: Firstly, this is in accordance with the standard practice followed in earlier tagsets. Secondly, grouping MWEs into a single unit should ideally be handled in chunking.

**Form vs. function:** We try to adopt a balance between form and function in a systematic and consistent way through deep analysis. Based on our analysis we propose to consider the *form* in normal circumstances and the *function* for words that are derived from other words. More details on this will be provided in the framework document (Baskaran et al 2007)

**Theoretical neutrality:** As Leech (1997) points out the annotation scheme should be theoretically neutral to make it clearly understandable to a larger group and for wider applicability.

**Diverse Language families:** As mentioned earlier, we consider eight languages coming from two major language families of India, viz. Indo-Aryan and Dravidian. Despite the distinct characteristics of these two families, it is however striking to note the typological parallels between them, especially in syntax. For example, both families follow SOV pattern. Also, several Indo-Aryan languages such as Marathi, Bangla etc. exhibit some agglutination, though not to the same extent of Dravidian. Given the strong commonalities between the two families we decided to use a single framework for them

#### 4 POS Tagset Framework for Indian languages

The tagset framework is laid out at the following four levels similar to EAGLES.

I. **Obligatory** attributes or values are generally universal for all languages and hence must be included in any morphosyntactic tagset. The major POS categories are included here.

II. **Recommended** attributes or values are recognised to be important sub-categories and features common to a majority of languages.

#### III. Special extensions<sup>1</sup>

- a. *Generic attributes* or values
- b. *Language-specific* attributes or values are the attributes that are relevant only for few languages and do not apply to most languages.

All the tags were discussed and debated in detail by a group of linguists and computer scientists/NLP experts for eight Indian languages viz. Bengali, Hindi, Kannada, Malayalam, Marathi, Sanskrit, Tamil and Telugu.

Now, because of space constraints we present only the partial tagset framework. This is just to illustrate the nature of the framework and the complete version as well as the rationale for different categories/features in the framework can be found in Baskaran et al. (2007).<sup>2</sup>

In the top level the following 12 categories are identified as universal categories for all ILs and hence these are obligatory for any tagset.

- |                     |                                |
|---------------------|--------------------------------|
| 1. [N] Nouns        | 7. [PP] Postpositions          |
| 2. [V] Verbs        | 8. [DM] Demonstratives         |
| 3. [PR] Pronouns    | 9. [QT] Quantifiers            |
| 4. [JJ] Adjectives  | 10. [RP] Particles             |
| 5. [RB] Adverbs     | 11. [PU] Punctuations          |
| 6. [PL] Participles | 12. [RD] Residual <sup>3</sup> |

The partial tagset illustrated in Figure 1 highlights entries in *recommended* and *optional* categories for verbs and participles marked for three levels.<sup>4</sup> The features take the form of attribute-value pairs with values in italics and in some cases (such as case-markers for participles) not all the values are fully listed in the figure.

#### 5 Current Status and Future Work

In the preceding sections we presented a common framework being designed for POS tagsets for Indian Languages. This hierarchical framework has

<sup>1</sup> We do not have many features defined under the special extensions and this is mainly retained for any future needs.

<sup>2</sup> Currently this is just the draft version and the final version will be made available soon

<sup>3</sup> For words or segments in the text occurring outside the gambit of grammatical categories like foreign words, symbols, etc.

<sup>4</sup> These are not finalised as yet and there might be some changes in the final version of the framework.

three levels to permit flexibility and interoperability between languages. We are currently involved in a thorough review of the present framework by using it to design the tagset for specific Indian languages. The issues that come up during this process will help refine and consolidate the framework further. In the future, annotation guidelines with some recommendations for handling ambiguous categories will also be defined. With the common framework in place, it is hoped that researchers working with Indian Languages would be able to not only reuse data annotated by each other but also share tools across projects and languages.

### References

Baskaran S. et al. 2007. Framework for a Common Parts-of-Speech Tagset for Indic Languages. (Draft) <http://research.microsoft.com/~baskaran/POSTagset/>

Cloeren, J. 1999. Tagsets. In *Syntactic Wordclass Tagging*, ed. Hans van Halteren, Dordrecht.: Kluwer Academic.

Hardie, A. 2004. The Computational Analysis of Morphosyntactic Categories in Urdu. PhD thesis submitted to Lancaster University.

Greene, B.B. and Rubin, G.M. 1981. Automatic grammatical tagging of English. Providence, R.I.: Department of Linguistics, Brown University

Garside, R. 1987 The CLAWS word-tagging system. In *The Computational Analysis of English*, ed. Garside, Leech and Sampson, London: Longman.

Leech, G and Wilson, A. 1996. Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R.

Leech, G. 1997. Grammatical Tagging. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, ed: Garside, Leech and McEnery, London: Longman

Leech, G and Wilson, A. 1999. Standards for Tag-sets. In *Syntactic Wordclass Tagging*, ed. Hans van Halteren, Dordrecht: Kluwer Academic.

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania

IIIT-tagset. A Parts-of-Speech tagset for Indian languages. [http://shiva.iiit.ac.in/SPSAL2007/iiit\\_tagset\\_guidelines.pdf](http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf)

Tamil tagset. AU-KBC Parts-of-Speech tagset for Tamil. [http://nrcfosshelpline.in/smedia/images/downloads/Tamil\\_Tagset-opensource.odt](http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt)

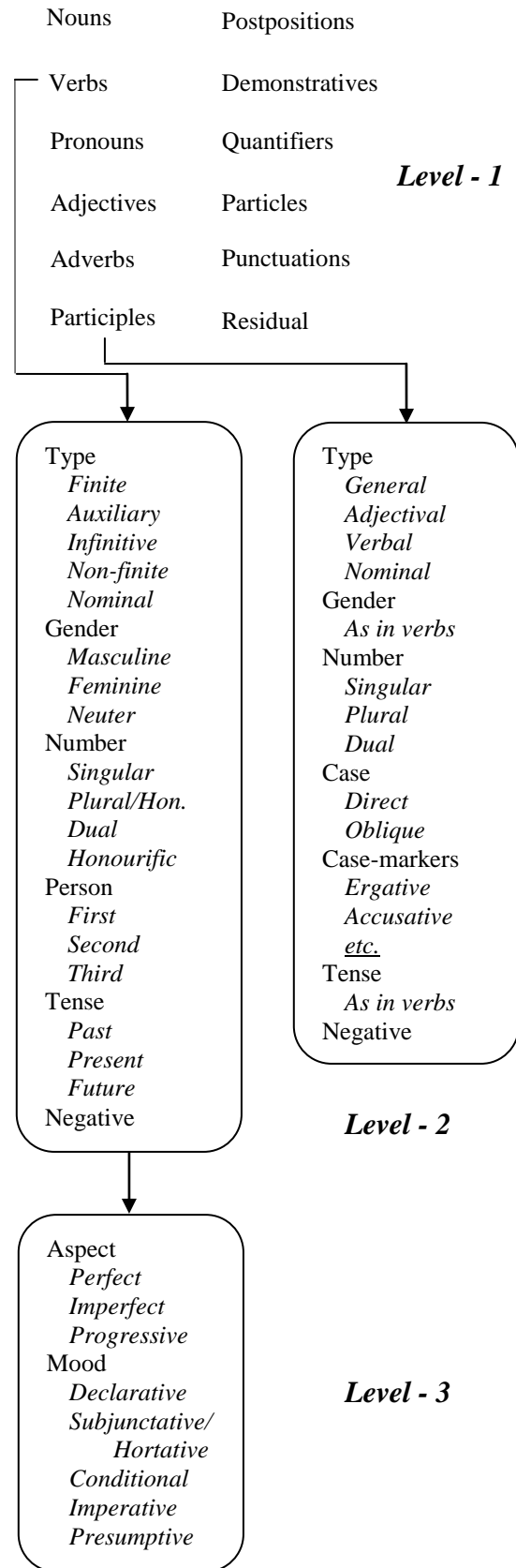


Fig-1. Tagset framework - partial representation

# Resources Report on Languages of Indonesia

**Hammam Riza**

IPTEKNET

Agency for the Assessment and  
Application of Technology (BPPT)

Jakarta, Indonesia

hammam@iptek.net.id

## Abstract

In this paper, we report a survey of language resources in Indonesia, primarily of indigenous languages. We look at the official Indonesian language (Bahasa Indonesia) and 726 regional languages of Indonesia (Bahasa Nusantara) and list all the available LRs that we can gathered. This paper suggests that the smaller regional languages may remain relatively unstudied, and unknown, but they are still worthy of our attention. Various LRs of these endangered languages are being built and collected by regional language centers for study and its preservation. We will also briefly report its presence on the Internet.

## 1 Introduction

It is not hard to get a picture of just how linguistically diverse Indonesia is. There are 726 languages in the country; making it the world's second most diverse, after Papua New Guinea which has 823 local languages (Martí et al., 2005:48).

The languages of Indonesia are part of a complex linguistic situation that is generally seen as comprised of three categories: Indonesian language, the regional indigenous languages, and foreign languages. (Alwi and Sugono, 2000).

The indigenous languages of Indonesia - also referred to as vernaculars or provincial languages, collectively called as Bahasa Nusantara - exhibits great variation in numbers of speakers. Thirteen of them have a million or more speakers, accounting

for 69.91% of the total population – Javanese (75,200,000 speakers), Sundanese (27,000,000), Malay (20,000,000), Madurese (13,694,000), Minangkabau (6,500,000), Batak (5,150,000), Buginese (4,000,000), Balinese (3,800,000), Acehese (3,000,000), Sasak (2,100,000), Makasarese (1,600,000), Lampungese (1,500,000), and Rejang (1,000,000). (Lauder, 2004: 3-4). Of these 13 languages, only 7 languages have presence on the Internet (Riza 2006).

The remaining 713 languages have a total population of only 41.4 million speakers, and the majority of these have very small numbers of speakers. For example, 386 languages are spoken by 5,000 or less; 233 have 1,000 speakers or less; 169 languages have 500 speakers or less; and 52 have 100 or less (Gordon, 2005). These languages are facing various degrees of language endangerment (Crystal, 2000).

There is evidence from census data over three decades that the growth in the numbers of speakers of Indonesian is reducing the numbers of speakers of the indigenous languages (Lauder, 2005). Concerns that this kind of growth would give Indonesian the potential to replace the regional languages were aired as early as the 1980s. (Poedjosoedarmo, 1981; Alisjahbana, 1984).

## 2 Language Resources

Many language centers in Indonesia have embarked in various research and development in creation of language resources (LRs). Unfortunately, this development mainly only focused on creating LRs for the official language Bahasa Indonesia. In the followings, we describe the present

and ongoing LRs research projects with emphasis on the indigenous languages.

### 2.1 Indonesian Electronic Dictionary System (KEBI)

Our laboratory has worked to enlarge and improve the quality of Indonesian electronic dictionaries. Starting 1987, it took us at least 4 years to develop all necessary components for CICC-MMTS, resulting with many first-ever Indonesian language resources, primarily electronic dictionaries and grammar rules for language analysis and generation. This extension have resulted in a collection of 500,000 word entries and more than 2 million derivational and inflected words. As part of this research, we built an online access to the dictionaries (<http://nlp.inn.bppt.go.id/kebi>) enabling users to add new words and definition. KBI electronic dictionary is scheduled to be launched in 2008, during 100 years celebration of the official Bahasa Indonesia.

### 2.2 BPPT-ANTARA Corpus

This parallel corpus was developed as extension to Indonesia National Corpus Initiative (INCI) which was earlier created to support the development of a hybrid stochastic-symbolic system BIAS-II. Currently, a pure statistical MT system based on Pharaoh is developed by BPPT and National News Agency (ANTARA) using 500K sentences pair, expected to have better accuracy and robustness and could enhance the quality of translation (current BLEU score 0.72).

### 2.3 Regional Languages Mapping (National Language Center)

For the past 15 years, the Indonesia National Language Center have been collecting information regarding all indigenous languages. By the end of this year, this project will be completed and all result and findings will be open to public.

### 2.4 Dictionaries of Bahasa Nusantara, Indonesian Linguistics Association (MLI)

Masyarakat Linguistik Indonesia (MLI) is a group of institutions, organizations and corporation, working together on mutually defined goals and projects that seek to provide a specification of LRs of all languages of Indonesia. MLI also help members to use the specification for tools and applica-

tions; find the best means to disseminate the specifications, tools and applications and encourage an open standard-based approach to the creation and interchange of LRs. It also demonstrate how MLI can be applied to Asian Language Resource (ALR) through making the results of collaborative endeavors available throughout the members of the group and wider associations; provide training, awareness and educational events and share with each other their work on related issues.

### 2.5 Speech Corpus

In a mission to improve the quality of automatic speech recognition (ASR), a collaboration of Telkom RDC and ATR-Japan has constructed speakers' corpus (40 speakers, 2000 sentences) which is expected to improve the accuracy of ASR to 90% level.

### 2.6 Other Corpus

Other monolingual corpus is found online. The major news articles corpora on the web is Tempointerakif.com (56,471 articles). Kompas corpus (71,109 articles) can be found at <http://ilps.science.uva.nl/Resources/BI>.

## References

- Alisjahbana, S. T. 1984. The problem of minority languages in the overall linguistic problems of our time. In *Linguistic Minorities and Literacy: Language Policy Issues in Developing Countries*, ed. F. Coulmas. Berlin: Mouton.
- Alwi, Hasan, and Sugono, Dendy. 2000. From National Language Politics to National Language Policy. Proceedings of the Seminar on Language Politics, Jakarta
- Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.
- Lauder, Multamia RMT. 2005. Language Treasures in Indonesia. In *Words and Worlds : World Languages Review*, eds. Fèlix Martí et al., 95-97. Clevedon [England] ; Buffalo [N.Y.]: Multilingual Matters.
- Martí, Fèlix, et.al. eds. 2005. *Words and Worlds : World Languages Review*. vol. 52. Bilingual Education and Bilingualism. Clevedon [England] ; Buffalo [N.Y.]: Multilingual Matters.
- Riza, H, et. al. 2006. Indonesian Languages Diversity on the Internet, Internet Governance Forum (IGF), Athens.

# Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List

Ryo Nishimura Yasuhiko Watanabe Yoshihiro Okada

Ryukoku University, Seta, Otsu, Shiga, 520-2194, Japan

r\_nishimura@afc.ryukoku.ac.jp

watanabe@rins.ryukoku.ac.jp

## Abstract

In this paper, we report a Japanese language resource for answering how-type questions. It was developed by using mails posted to a mailing list. We show a QA system based on this language resource.

## 1 Introduction

In this paper, we report a Japanese language resource for answering how type questions. It was developed by using mails posted to a mailing list and it was given the four types of descriptions: (1) mail type, (2) key sentence, (3) semantic label, and (4) credibility label. Credibility is a center problem of knowledge acquisition from natural language documents because the documents, including mails posted to mailing lists, often contain incorrect information. We describe how to develop this language resource in section 2, and show a QA system based on it in section 3.

## 2 Language resource development

There are mailing lists to which question and answer mails are posted frequently. For example, to Vine Users ML, considerable number of question mails and their answer mails are posted by participants who are interested in Vine Linux <sup>1</sup>. We intended to use these mails for developing a language resource because we have the following advantages.

- It is easy to collect question and answer mails in various domains: The sets of question and answer mails are necessary to answer how-type questions. Many informative mails posted to mailing lists are disclosed in the Internet and can be retrieved by using full text search engines, such as Namazu (Namazu). However, users want a more convenient retrieval system than existing systems.
- There are many mails which report the credibility of their previous mails: Answer mails often contain incorrect solutions. On the other hand, many

<sup>1</sup>Vine Linux is a linux distribution with a customized Japanese environment.

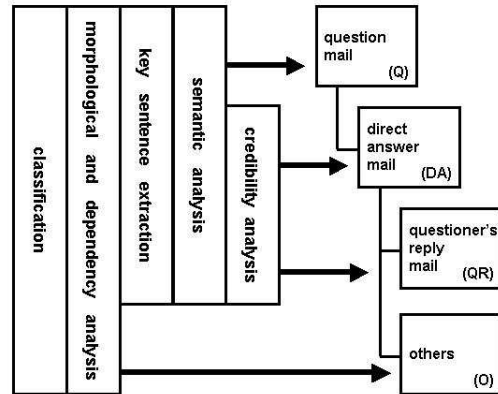


Figure 1: The overview of the language resource development

mails were submitted by questioners for reporting the credibility of the solutions which they had received. As a result, solutions described in answer mails can be confirmed by using questioner's reply mails.

- Mails posted to mailing lists generally have key sentences: These key sentences can be extracted by using surface clues (Watanabe 05). The sets of questions and solutions can be acquired by using key sentences in question and answer mails. Also, the solutions are confirmed by using key sentences in questioner's reply mails. Furthermore, key sentences in question mails and their neighboring sentences often contain information about conditions, symptoms, and purpose. These kinds of information are useful in specifying user's unclear questions.

Figure 1 shows the overview of the language resource development. First, by using reference relations and sender's email address, mails are classified into four types: (1) question (Q) mail, (2) direct answer (DA) mail, (3) questioner's reply (QR) mail, and (4) others. DA mails are direct answers to the original questions. Solutions are generally described in the DA mails. QR mails are questioners' answers to the DA mails. In the QR mails, questioners often report the

credibility of the solutions described in the DA mails. Sentences in the Q, DA, and QR mails are transformed into dependency trees by using JUMAN(JMN 05) and KNP(KNP 05).

Second, key sentences are extracted from the Q, DA, and QR mails by using (1) nouns used in the mail subjects, (2) quotation frequency, (3) clue expressions, and (4) sentence location (Watanabe 05). To evaluate this method, we selected 100 examples of question mails and their DA and QR mails in Vine Users ML. The accuracy of the key sentence extraction from the Q, DA, and QR mails were 80%, 88%, and 76%, respectively. We associated (1) the key sentences and the neighboring sentences in the Q mails and (2) the key sentences in the DA mails. We used them as knowledge for answering how-type questions. 73% of them were coherent explanations.

Third, expressions including information about condition, symptom, and purpose are extracted from the key sentences in the Q mails and their neighboring sentences by using clue expressions. The results are used for specifying unclear questions. For example, unclear question “*oto ga denai* (I cannot get any sounds)” is specified by “*saisho kara* (symptom: from the beginning) ?” and “*kernel no version ha* (condition: which kernel version) ?”, both of which were extracted from the Q mails through this semantic analysis. The accuracy of this analysis was 74%.

Finally, positive and negative expressions to the solutions described in the DA mails are extracted from the key sentences in the QR mails. The results of this analysis on QR mails are used for giving credibility labels to the solutions described in the DA mails. The accuracy of this analysis was 76%.

### 3 QA system based on the language resource

Figure 2 shows the overview of our system based on the language resource. A user can ask a question to the system in natural language. Then, the system retrieves similar questions and their solutions, and it shows the credibility of these solutions by using their credibility labels. Figure 3 shows an example where our system gave an answer to user’s question, “*IP wo shutoku dekinai* (I cannot get an IP address)”; “positive 1” means that this answer thread has one solution that was positively confirmed by its QR mail.

The language resource consists of the mails posted to Vine Users ML (50846 mails: 8782 Q mails, 13081

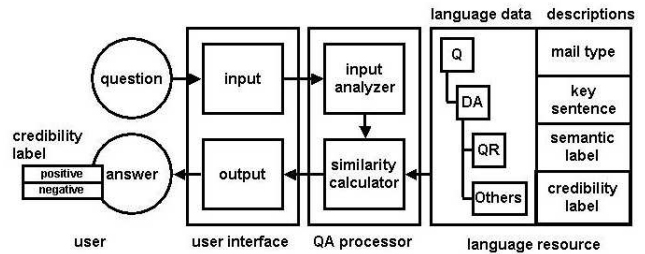


Figure 2: System overview

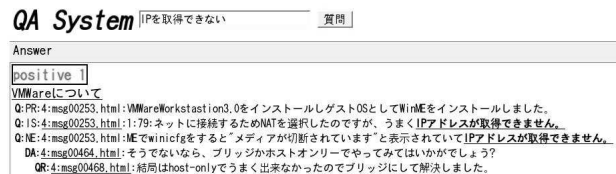


Figure 3: A set of a question and the answers with a positive label retrieved by our system

DA mails, 4272 QR mails, and 24711 others). 8782 key sentences and their 7330 previous and 8614 next sentences were extracted from the Q mails. These sentences were associated with 13081 key sentences extracted from the DA mails and used as knowledge for answering how-type questions. 3173 key sentences were extracted from the QR mails and the credibility labels (2148 positive and 1025 negative) were given to 3127 key sentences in the DA mails.

The QA processor transforms user’s question into a dependency structure by using JUMAN(JMN 05) and KNP(KNP 05). Then, it retrieves similar questions and their solutions by calculating the similarity scores between user’s question and key sentences in the question mails. It also retrieves expressions including information about conditions, symptoms, and purpose which seem to be useful in specifying user’s questions.

The user interface enables a user to access to the system via a WWW browser by using CGI-based HTML forms. It puts the answer threads in order of similarity score.

### References

- Namazu: a Full-Text Search Engine, <http://www.namazu.org/>
- Watanabe, Nishimura, and Okada: Confirmed Knowledge Acquisition Using Mails Posted to a Mailing List, IJCNLP 2005, pp.131-142, (2005).
- Kurohashi and Kawahara: JUMAN Manual version 5.1 (in Japanese), Kyoto University, (2005).
- Kurohashi and Kawahara: KNP Manual version 2.0 (in Japanese), Kyoto University, (2005).



# Corpus Building for Mongolian Language

**Purev Jaimai**

Center for Research on Language Processing,  
National University of Mongolia, Mongolia  
purev@num.edu.mn

**Odbayar Chimeddorj**

Center for Research on Language Processing,  
National University of Mongolia, Mongolia  
odbayar@num.edu.mn

## Abstract

This paper presents an ongoing research aimed to build the first corpus, 5 million words, for Mongolian language by focusing on annotating and tagging corpus texts according to TEI XML (McQueen, 2004) format. Also, a tool, MCBUILDER, which provides support for flexibly and manually annotating and manipulating the corpus texts with XML structure, is presented.

## 1 Introduction

Mongolian researchers quite recently have begun to be involved in the research area of Natural Language Processing. All necessary linguistic resources, which are required for Mongolian language processing, have to be built from scratch, and then they should be shared in public research for the rapid development of Mongolian language processing.

This ongoing research aims to build a tagged and parsed 5 million words corpus for Mongolian by developing a spell-checker, tagger, sentence-parser and others (see Figure 1 and 2). Also, we needed to develop a tagset for the corpus because there was not any tagset for Mongolian and the traditional words categories are not appropriate to it. Thus, we designed a high level tagset, which consists of 20 tags, and are further classifying them. Currently, we have collected and populated 500 thousand words, 50 thousand of which have been manually tagged, into the corpus (see Figure 1).

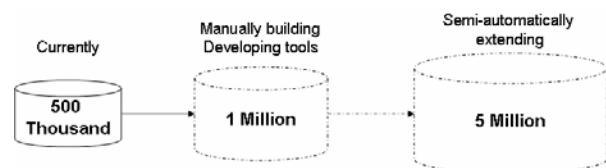


Figure 1. Current and future states of building a Mongolian corpus.

And, we manually build the corpus until collecting and annotating 1 million words and tagging 100 thousand words of them for semi-automatically building the corpus in the future.

## 2 Corpus Building Design

We are building the corpus as sub-corpora, which are a raw corpus, a cleaned corpus, a tagged corpus and a parsed corpus, separately for various kinds of studying and use on Mongolian language (Figure 2).

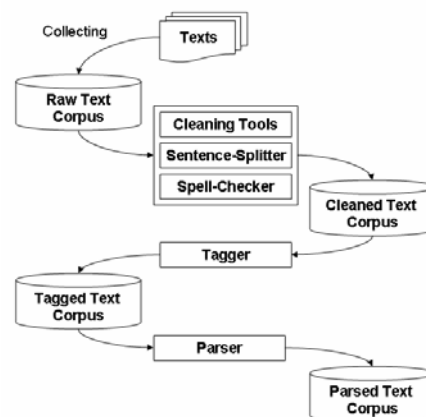


Figure 2. Schema of building a Mongolian corpus.

At first, we are collecting the editorial articles of Unen newspaper (Unen publish), which is one of

the best written newspapers in Mongolia, by using OCR application. We will also collect laws, school book, and literary text (see Figure 3).

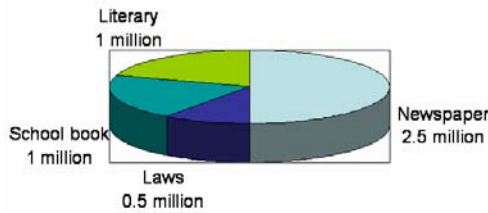


Figure 3. Text sizes included in the corpus.

The corpus annotation follows TEI XML standard. According to the work scope, the annotating part is divided into two parts that are structural annotation such as paragraphs, sentences, and so on, and POS tagging.

The structure of the text annotation is presented in Figure 4.

```

<tei>
  <teiHeader>
    <fileDesc />
  </teiHeader>
  <text>
    <body>
      <s>
        <word id=" " pos="tag">WORD</word>
      </s>
    </body>
  </text>
</tei>

```

Figure 4. XML Structure of corpus text.

For annotating two parts, once a manual corpus builder, called MCBuilder, were planned to develop, we have developed the first version and used to annotating 500 thousand word texts and tagging 50 of them (see Figure 5).

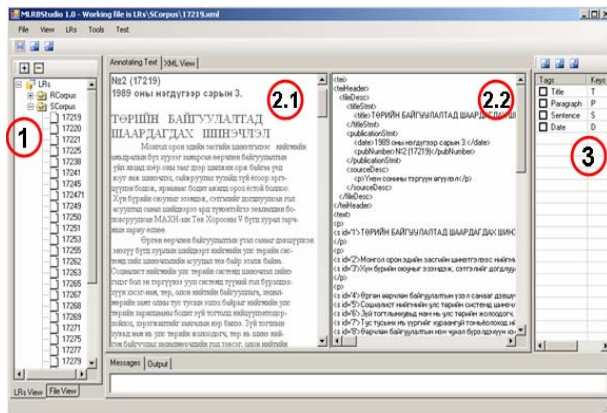


Figure 5. Screenshot of the corpus organizer and its main view.

MCBuilder has three main windows that are (1) manipulating and organizing the corpus, (2) annotating sample texts and (3) manipulating tagset as shown in Figure 5.

### 3 Conclusion

Mongolian language has hardly studied by computer, and its traditional rules such as inflectional, derivational, part of speech, sentence constituents, etc are extremely difficult to computerize. Our research works in the last few years showed it (Purev, 2006). Therefore, we are revising them by creating a corpus for computer processing.

The proposals of this ongoing research are the first Mongolian 5 million words corpus, and tools that are spell-checker, tagger and parser.

Currently, we have done followings:

- Defined the corpus design, XML structure of the corpus text, and the high level tagset
- Collected and annotated 500 thousand words text
- Tagged 50 thousand words
- Released the first version of a Mongolian corpus building tool called MCBuilder
- First versions of Syllable-parser and Morph-analyzer for Mongolian

We are planning to complete the corpus in the next two years.

### 4 Acknowledgement

Here described work was carried out by support of PAN Localization Project (PANL10n).

### References

PANL10n: PANLocalization Project. National University of Computer and Emerging Sciences, Pakistan.

Purev J. 2006. *Corpus for Mongolian Language*, Research Project, Mongolia.

Purev J. and Odbayar Ch.. 2006. *Towards Constructing the Corpus of Mongolian Language*, Proceeding of ICEIC.

Sperberg-McQueen, C. M. and Burnard, L.. 2004. *Text Encoding Initiative. The XML version of the TEI Guidelines*, Website.

Unen press. 1984-1989. *Editorial Articles*. Mongolia

# Resources for Urdu Language Processing

Sarmad Hussain

Center for Research in Urdu Language Processing  
National University of Computer and Emerging Sciences  
B Block, Faisal Town, Lahore, Pakistan  
sarmad.hussain@nu.edu.pk

## Abstract

Urdu is spoken by more than 100 million speakers. This paper summarizes the corpus and lexical resources being developed for Urdu by the CRULP, in Pakistan.

## 1 Introduction

Urdu is the national language of Pakistan and one of the state languages of India and has more than 60 million first language speakers and more than 100 million total speakers in more than 20 countries (Gordon 2005). Urdu is written in Nastalique writing style based on Perso-Arabic script. This paper focuses on the Urdu resources being developed, which can be used for research in computational linguistics.

## 2 Urdu Text Encoding

Urdu computing started early, in 1980s, creating multiple encodings, as a standard encoding scheme was missing at that time. With the advent of Unicode in early 1990s, some online publications have switched to Unicode, but much of the publication still continues to follow the ad hoc encodings (Hussain et al. 2006). Two main on-line sources of Urdu text in Unicode are Jang News ([www.Jang.net/Urdu](http://www.Jang.net/Urdu)) and BBC Urdu service ([www.BBC.co.uk/Urdu](http://www.BBC.co.uk/Urdu)) and are thus good sources of corpus. Encoding conversion may be required if data is acquired from other sources.

## 3 Corpora

EMILLE Project, initiated by Lancaster University is one of the first initiatives to make Urdu corpus available for research and development of language processing (McEnery et al. 2000). The project has released 200,000 words of English text translated into Bengali, Gujarati, Hindi, Punjabi

and Urdu, creating a parallel corpus across these languages. In addition, the corpus also has 512,000 words of Spoken Urdu, from BBC Radio. Moreover, the corpus also contains 1,640,000 words of Urdu text. These Urdu corpus resources are also annotated with a large morpho-syntactic tag-set (Hardie 2003).

Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan has also been developing corpora and associated tools for Urdu. A recent project collected a raw corpus of 19 million words of Urdu text mostly from Jang News, reduced to 18 million words after cleaning. The corpus collection has been based on LC-STAR II guidelines<sup>1</sup>. The domain-wise figures are given in Table 1. Further details of the corpus and associated information are discussed by Ijaz et al. (2007).

Table 1: Distribution of Urdu Corpus

Domains	Cleaned Corpus	
	Total Words	Distinct Words
C1. Sports/Games	1529066	15354
C2. News	8425990	36009
C3. Finance	1123787	13349
C4. Culture/Entertainment	3667688	34221
C5. Consumer Information	1929732	24722
C6. Personal communications	1632353	23409
Total	18308616	50365

Agreement between CRULP and Jang News allows internal use. However, due to distribution restrictions in this agreement, the corpus has not been made publicly available. The distribution rights are still being negotiated with Jang News.

The tag set developed by Hardie (2003) is based on morpho-syntactic analysis. A (much reduced) syntactic tag set has also been developed by

<sup>1</sup> See [www.lc-star.org/docs/LC-STAR\\_D1.1\\_v1.3.doc](http://www.lc-star.org/docs/LC-STAR_D1.1_v1.3.doc)

CRULP (on the lines of PENN Treebank tagset), available at its website [www.CRULP.org](http://www.CRULP.org). A corpus of 100,000 words manually tagged on this tag set has also been developed based on text from Jang online news service. This *CRULP POS Tagged Jang News Corpus* is available through the center.

Recently another corpus of about 40,000 words annotated with Named Entity tags was also made available for Workshop on NER for South and South East Asian Languages organized at IJCNLP 2008. The annotated corpus was donated by CRULP and IIIT Hyderabad and is available at <http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5>.

Tag set contains 12 tags. Details of these tags are discussed at the link <http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=3>. The CRULP portion of the data is also available at CRULP website, and is a subset of the *CRULP POS Tagged Jang News Corpus*.

In earlier work at CRULP, a 230 spelling errors corpus has also been developed based on typographical errors in Newspapers and student term papers. See Naseem et al. (2007) for details.

A corpus of Urdu Names has also been developed by CRULP, based on the collective telephone directories of Pakistan Telecommunications Corporation Limited (PTCL) from across all major cities of Pakistan. A name list has also been extracted from the corpus for all person names, addresses and cities of Pakistan.

## 4 Lexica

Lexica are as critical for development of language computing as corpora. One of the most comprehensive lexica available for Urdu was recently released by CRULP (available through CRULP website). The online version, called Online Urdu Dictionary (OUD) contains 120,000 entries, with 80,000 words annotated with significant information. The data of OUD is XML tagged, as per the annotation schema discussed by Rahman (2005; pp. 15), which contains about 20 etymological, phonetic, morphological, syntactic, semantic and other parameters of information about a word. The dictionary also gives translation of 12000 words in English and work is under way to enable runtime user-defined queries on the available XML tags. The contents of this lexicon are based on the 21 volume Urdu Lughat

developed by Urdu Dictionary Board of Government of Pakistan. See [www.crupl.org/oud](http://www.crupl.org/oud) for details.

CRULP has also developed a corpus based lexicon of 50,000 words with frequency data and annotation specifications defined by LC-STAR II project (at [http://www.lc-star.org/docs/LC-STAR\\_D1.1\\_v1.3.doc](http://www.lc-star.org/docs/LC-STAR_D1.1_v1.3.doc)). Details of the lexicon annotation scheme are given by Ijaz et al. (2007).

There are also additional tools available through CRULP, and documented at its website, including normalization, collations, spell checking, POS tagging and word segmentation applications.

## 5 Conclusions

This paper lists some core linguistic resources of Urdu, available through CRULP and other sources. However, the paper identifies licensing constraints, a challenge for open distribution, which needs to be addressed.

## References

- Gordon, Raymond G., Jr. (ed.). (2005). *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com/>.
- Hardie, A. (2003). Developing a tag-set for automated part-of-speech tagging in Urdu. In Archer, D, Rayson, P, Wilson, A, and McEnery, T (eds.) *Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16*. Department of Linguistics, Lancaster University, UK.
- Ijaz, M. and Hussain, S. (2007). Corpus Based Urdu Lexicon Development. In the *Proceedings of Conference on Language Technology '07*, University of Peshawar, Peshawar, Pakistan.
- Naseem, T. and Hussain, S. (2007). Spelling Error Trends in Urdu. In the *Proceedings of Conference on Language Technology '07*, University of Peshawar, Peshawar, Pakistan.
- McEnery, A., Baker, J., Gaizauskas, R. & Cunningham, H. (2000). EMILLE: towards a corpus of South Asian languages, *British Computing Society Machine Translation Specialist Group*, London, UK.
- Rahman, S. (2005). Lexical Content and Design Case Study. Presented at *From Localization to Language Processing, Second Regional Training of PAN Localization Project*. Online presentation version: <http://pan10n.net/Presentations/Cambodia/Shafiq/LexicalContent&Design.pdf>.

# Balanced Corpus of Contemporary Written Japanese

**Kikuo Maekawa**

Dept. Language Research, National Institute for Japanese Language  
10-2, Midori-cho, Tachikawa-shi, Tokyo 190-8561 JAPAN  
kikuo@kokken.go.jp

## Abstract

Construction of 100 million words balanced corpus of contemporary written Japanese is underway at the National Institute for Japanese Language. The unique property of the corpus consists in that the majority of its sample texts are selected randomly from well-defined statistical populations covering wide range of written texts.

## 1 Introduction

A serious problem in corpus-based analysis of the Japanese language is the lack of reliable balanced corpus. Most of the corpus-based studies of contemporary Japanese are based upon the analyses of text archive of newspaper articles, archive of copyright-expired literary work (*Aozora Bunko*), or crawling of the Internet text.

Putting aside the problems of the copyright-expired literary works, which are definitely too old to be the material for the study of the contemporary Japanese, lack of balanced corpus imposes two mutually related problems on linguistic studies. Most of newspaper articles are written by newspaper writers who are very much aware of the established writing style (and orthography) of newspaper articles. Accordingly, it is the genre of text where variations of all sorts are suppressed to the minimum level.

On the other hand, the results of internet crawling using search engines like Google are very much likely to include texts covering wide range of texts. It is also expected that considerable amount of linguistic variations are to be observed.

It is, however, very difficult, if not impossible, to conduct analyses of style difference and/or lin-

guistic variations using the results of internet crawling, because the information about the genre of texts and/or the writers are usually missing. Moreover, the amount of retrieved texts can often be too large to be classified by hand.

There is also a problem of skewed distribution of texts caused by copyright protection. Copyright-protected materials, especially literary works, will not usually show up in the Internet.

To solve these problems in Japanese corpus linguistics, National Institute for Japanese Language (NIJL, hereafter) has launched a corpus compilation project in the spring of 2006, aiming at public release of Japan's first 100 million words balanced corpus in the year of 2011. The corpus is named the *Balanced Corpus of Contemporary Written Japanese*, or BCCWJ.

<b>PUBLICATION (PRODUCTION) SUB-CORPUS</b> Books, Magazines, and Newspapers, 35 million words. 2001-2005	<b>LIBRARY (CIRCULATION) SUB-CORPUS</b> Books 30 million words. 1986-2005
<b>SPECIAL PURPOSE SUB-CORPUS</b> Whitepaper, Diet minute, Web text, Textbooks, etc. 35 million words. 1975-2005	

Figure 1. The three components of the BCCWJ.

## 2 Design of the BCCWJ

As shown in the figure 1, the BCCWJ consists of 3 component sub-corpora, viz., 'publication', 'library', and 'special purpose' sub-corpora.

### 2.1 Publication sub-corpus

The upper left-hand sub-corpus of the figure 1 is called 'publication' sub-corpus. This is also called

‘production’ sub-corpus. As the name suggests, this sub-corpus represents the production, as opposed to the reception aspect of contemporary written Japanese. The sub-corpus consists of samples extracted randomly from the statistical population covering the whole body of books, magazines, and newspapers published during 2001-2005.

The population was constructed using the sources that are publicly available; *J-BISC* (Japan Biblio Disc) and *Periodicals in Print in Japan* were used as the sources for books and magazines respectively. The data for newspapers was available from the association of newspaper companies (*Nihon Shinbun Kyokai*). The total number of characters involved in the population was estimated and samples for the BCCWJ were drawn in the way that each character in the population had the same chance of being sampled.

It is to be noted at this point that the composition ratios among text genres (i.e. the ratio among samples of books, magazines, and newspapers) were determined on the basis of publicly available data mentioned above. This makes crucial difference from the designs of corpora like the Brown Corpus and the BNC, where the composition ratios of various genres were determined subjectively by specialists of the English language without making reference to any objective data.

The total size of the sub-corpus is supposed to be about 34.7 million words; and, 74.1, 16.1, and 9.8% of the sub-corpus are to be devoted to the samples of books, magazines, and newspapers respectively.

## 2.2 Library sub-corpus

The second sub-corpus is called ‘library’ or ‘circulation’ sub-corpus. The sampling population for this sub-corpus was the whole books registered in at least 13 public libraries in the Tokyo Metropolis. The population thus defined contains about 335,000 books. According to our estimation, more than 48 billion characters are included in this population, which is nearly the same amount as the population of the book part of the publication sub-corpus.

There are two important differences between the publication and library sub-corpora. Firstly, the texts of the library sub-corpora represent those books that were accepted by a certain number of readers, while the texts in the publication sub-corpora have no guarantee of the sort.

Secondly, the library sub-corpus covers the period of time 1986-2005 (1986 was the year when the ISBN book classification system was adopted by most of major publishing companies), while the period of time covered by the publication sub-corpus is 2001-2005.

## 2.3 Special-purpose sub-corpus

The last sub-corpus is called ‘special purpose’ or ‘out-of-population’ sub-corpus. This is the aggregate of various special purpose mini corpora, and, unlike the former two sub-corpora, some of the mini corpora are not sampled using the technique of random sampling (because the populations can not be defined).

The mini corpora currently include texts of governmental white papers, Internet text (Yahoo! Japan’s bulletin board *Chiebukuro*), minutes of the national diet, school textbooks, and best-selling books of the past 30 years. Laws and academic papers will also be included.

Most of these mini corpora contain about 5 million words, and will be utilized in the language policy oriented activities of the NIJL, including the revision of the National list of Chinese Characters for Daily Usage (*Jouyou kanji hyou*).

## 3 Funds and the current status

The compilation of the BCCWJ is supported by the budget of the NIJL and the MEXT (ministry of education) Grant-in-Aid for Scientific Research Priority Area Program “*Japanese Corpus*” (2006-2010) [1].

As of September 2007, texts containing about 30 million words have been sampled and stored in the NIJL server, and, the full-text query of the 10 million words texts that are copyright cleared are publicly available on the web [2].

Upon its completion, the BCCWJ will be the world’s first balanced corpus that is designed and compiled based upon rigid statistical sampling. This will open up a new possibility in Japanese linguistics, and the design of language corpora in general.

## References

- [1] <http://www.tokuteicorpus.jp/>
- [2] <http://www.kotonoha.gr.jp/demo/>

# A Basic framework to Build a Test Collection for the Vietnamese Text Categorization

Viet Hoang-Anh, Thu Dinh-Thi-Phuong, Thang Huynh-Quyet

Hanoi University of Technology, Vietnam

vietha-fit@mail.hut.edu.vn, thanghq-fit@mail.hut.edu.vn

## Abstract

The aim of this paper is to present a basic framework to build a test collection for a Vietnamese text categorization. The presented content includes our evaluations of some popular text categorization test collections, our researches on the requirements, the proposed model and the techniques to build the BKTexts - test collection for a Vietnamese text categorization. The XML specification of both text and metadata of Vietnamese documents in the BKTexts also is presented. Our BKTexts test collection is built with the XML specification and currently has more than 17100 Vietnamese text documents collected from e-newspapers.

## 1 Introduction

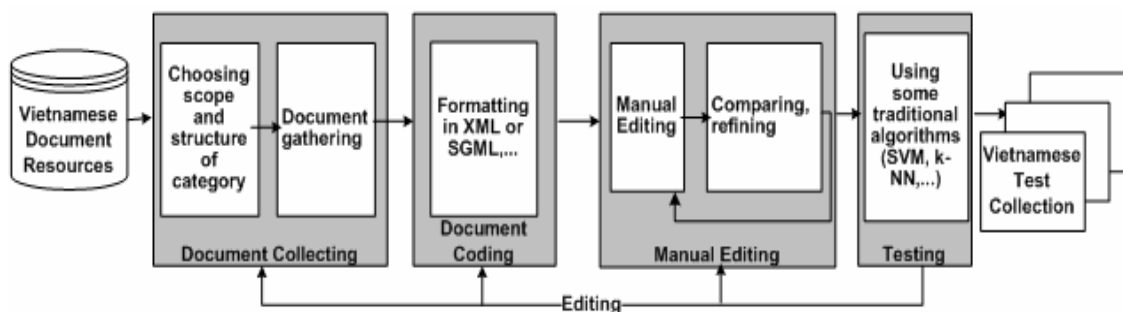


Figure1. The system architecture to build a Vietnamese test collection for text categorization

Natural Language Processing (NLP) for such popular languages as English, French, etc. has been well studied with many achievements. In contrast, NLP for unpopular languages, such as Vietnamese, has only been researched recently. It means that expecting international scientists to care about our problems is not feasible in the near future. In this paper, we present our research results on that field, especially on Vietnamese test collections for Vietnamese text categorization. This paper will be

organized as follows. Section 2 proposes our research on the requirements, models and techniques to build a Vietnamese test collection for researches and experiments on Vietnamese text categorization. Section 3 presents our results with BKTexts test collection. Lastly, the focus of our ongoing research will be presented in section.

## 2 Model of building a test collection for the Vietnamese text categorization

Until now, there has not been a Vietnamese standard test collection for Vietnamese text categorization. Vietnamese documents used in previous studies of Vietnamese researchers are gathered by themselves and were not thoroughly checked. Moreover, all over the world, there have been a lot of test collections in many different languages, especially in English such as the Reuters-21578, the RCV1 and the 20NewsGroup<sup>1</sup>. Therefore, we intend to build a Vietnamese standard test collection

for the Vietnamese text categorization. We defined a framework for building Vietnamese test collections as follows. Basic requirements for a Vietnamese test collection text categorization

Our model to build a Vietnamese test collection for text categorization is accomplished in four stages: collecting, auto coding, manual editing, and testing (Figure 1).

<sup>1</sup> Available from <http://kdd.ics.uci.edu/>

From available resources, we gather Vietnamese documents for the test collection in accordance with the scope and the structure of categories. Researchers usually use documents collected from e-newspapers because these documents are pre-processed and less ambiguous. Then an auto system tags documents in the XML (or SGML) formatting specification.

After being coded, documents are manual edited by editors. The editors would assign the categories they felt applicable. They also edit specification tags of formatted documents in order to completely and more precisely describe attributes of documents. Lastly, to assess the accuracy of the test collection, we use some famous categorization algorithms such as SVM, k-NN, etc. Performing the test and correction several times, we will gradually obtain a finer and more precise test collection. The process ends when errors are below a permitted threshold.

### 3 The BKTexts test collection for Vietnamese text categorization

With the model mentioned above, we are constructing the BKTexts test collection for the first version. We collected about 17100 documents for the BKTexts from two e-newspapers <http://www.vnexpress.net> and <http://www.vnn.vn>. Categories are organized in a hierarchical structure of 10 main classifications and 37 sub-classes. Documents are marked up with XML tags and given unique ID numbers. The XML specification of a document in the BKTexts test collection is described in Figure 2. Building a successful Vietnamese test collection for text categorization has a significant meaning. It will be a useful material for any study on text categorization and Vietnamese processing in the future because it reduces a lot of manual work and time, as well as increases the accuracy of experimental results.

### 4 Conclusion and future work

We have presented our research results on defining requirements, the model and techniques to build a Vietnamese test collection for researches and experiments on Vietnamese text categorization. Currently, we continue building the BKTexts on a larger scale for publishing widely in the near future. This test collection enables researchers to test ideas

and to objectively compare results with published studies.

```

<?xml version="1.0" encoding="Unicode" ?>
<BKTEXTS ID="" SPLIT=""> // Identifying the document ID
and whether it belongs to the training or test set.
<METADATA>
  <TOPICS></TOPICS> // Categories of the document
  <DATE></DATE> // the date of the document
  <VNFONT></VNFONT> // the Vietnamese font of the
document
  <SIZE></SIZE> // the size of the document
  <UNKNOWN_TEXT> </UNKNOWN_TEXT> // the noisy
characters in the document
  <SOURCE> // the source of the document
  <DATELINE></DATELINE>
  <ORGS></ORGS>
  <COUNTRIES></COUNTRIES>
  </SOURCE>
  <AUTHOR> // authors of the document
  <FULLNAME></FULLNAME>
  <ORGS></ORGS>
  <COUNTRIES></COUNTRIES>
  </AUTHOR>
  <CODER> // the editor coding the document
  <FULLNAME></FULLNAME>
  <ORGS></ORGS>
  <COUNTRIES></COUNTRIES>
  <NOTES></NOTES> // some notes of the editor
  </CODER>
</METADATA>
<TEXT>
  <TITLE></TITLE> // Title of the document
  <SUMMARY></SUMMARY> // Summary of the document
  <HEADLINE></HEADLINE> //
  <BODY></BODY> // the main text of the document
</TEXT>
<COPYRIGHT></COPYRIGHT>
</BKTEXTS>

```

Fig.2. The XML specification of the BKTexts

### References

- David D. Lewis, Reuters-21578 Text Categorization Test Collection, [www.daviddlewis.com](http://www.daviddlewis.com), 1997.
- David D. Lewis, Yiming Yang, Tony G.Rose, Fan Li, "RCV1: A new Benchmark Collection for Text Categorization Research", in: *Journal of Machine Learning Research* 5, pp.361-397, 2004.
- Huynh Quyet Thang, Dinh Thi Phuong Thu. Vietnamese text categorization based on unsupervised learning method and applying extended evaluating formulas for calculating the document similarity. *Proceedings of The Second Vietnam National Symposium on ICT.RDA, Hanoi 24-25/9/2004*, pp. 251-261 (in Vietnamese)
- Dinh Thi Phuong Thu, Hoang Vinh Son, Huynh Quyet Thang. Proposed modifications of the CYK algorithm for the Vietnamese parsing. *Journal of Computer Science and Cybernetics, Volume 21, No. 4, 2005*, pp. 323-336 (in Vietnamese)



# Enhanced Tools for Online Collaborative Language Resource Development

**Virach Sornlertlamvanich**  
**Thatsanee Charoenporn**  
**Suphanut Thayaboon**  
**Chumpol Mokrat**

Thai Computational Linguistics Lab.  
NICT Asia Research Center,  
Pathumthani, Thailand  
{virach, thatsanee, suphanut,  
chumpol}@tccllab.org

**Hitoshi Isahara**

National Institute of Information  
and Communications Technology  
3-5 Hikaridai, Seika-cho, soraku-gaun,  
Kyoto, Japan 619-0289  
isahara@nict.go.jp

## Abstract

This paper reports our recent work of tool development for language resource construction. To make a revision of Asian WordNet which is automatically generated by using the existing English translation dictionary, we propose an online collaborative tool which can organize multiple translations. To support the work of syntactic dependency tree annotation, we develop an editing suite which integrates the utilities for word segmentation, POS tagging and dependency tree into a sequence of editing.

## 1 Introduction

Though WordNet was already used as a starting resource for developing many language WordNets, the constructions of the WordNet for languages can be varied according to the availability of the language resources. Some were developed from scratch, and some were developed from the combination of various existing lexical resources.

This paper presents an online collaborative tool particularly to facilitate the construction of the Asian WordNet which is automatically generated by using the existing resources having only English equivalents and the lexical synonyms.

In addition, to support the work of syntactic dependency tree annotation, we develop an editing suite which integrates the utilities for word segmentation, POS tagging and dependency tree. The tool is organized in 4 steps, namely, sentence selection, word segmentation, POS tagging, and syntactic dependency tree annotation.

The rest of this paper is organized as follows: Section 2 describes the collaborative interface for revising the result of synset translation. Section 3 describes the tool for annotating Thai syntactic dependency tree corpus. And, Section 4 concludes our work.

## 2 Collaborative Tools for Asian WordNet Construction

There are some efforts in developing Wordnets for some of Asian languages, e.g. Chinese, Japanese, Korean, and Hindi. The number of languages that have been successfully developed their Wordnets is still limited to some active research in the area. However, the extensive development of Wordnet in other languages is of the efforts to support the NLP research and implementation. It is not only to facilitate the implementation of NLP applications for the language, but also provide an inter-linkage among the Wordnets for different languages to develop multi-lingual applications.

We adopt the proposed criteria for automatic synset assignment for Asian languages which has limited language resources. Based on the result from the above synset assignment algorithm, we introduce KUI (Knowledge Unifying Initiator), (Sornlertlamvanich et al., 2007) to establish an online collaborative work in refining the WordNets.

KUI is a community software tool which allows registered members including language experts to revise and vote for the synset assignment. The system manages the synset assignment according to the preferred score obtained from the revision process. As a result, the community-based WordNets will be accomplished and exported into the original form of WordNet database. Via the synset

ID assigned in the WordNet, the system can generate a cross language WordNet. Through this effort, a translated version of Asian WordNet can be established.

Table 1 shows a record of WordNet displayed for translation in KUI interface. English entry together with its part-of-speech, synset, and gloss are provided if exists. The members will examine the assigned lexical entry and decide whether to vote for it or propose a new translation.

Car
[Options]
POS : NOUN
Synset : auto, automobile, machine, motorcar
Gloss : a motor vehicle with four wheels; usually propelled by an internal combustion engine;

Table 1. A record for a synset

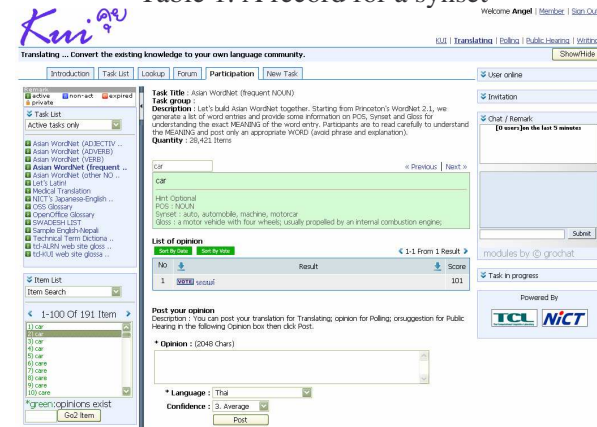


Figure 1. KUI interface (www.tcllab.org/kui)

Figure 1 illustrates the translation page of KUI. In the working area, the login member can participate in proposing a new translation or voting for the preferred translation to revise the synset assignment. Statistics of the progress as well as many useful functions such as item search, chat, and list of online participants are also provided to understand the progress of work and to work online with other members.

### 3 Tool for Constructing a Syntactic Dependency Tree Annotated Corpus

The tool is organized in 4 steps, namely, sentence selection, word segmentation, POS tagging, and syntactic dependency tree annotation, shown in Figure 2. Sentence segmentation is yet another crucial issue for the Thai language. We, however, will not discuss about the issue in this work. The input is already a list of sentences provided for annotator to select.

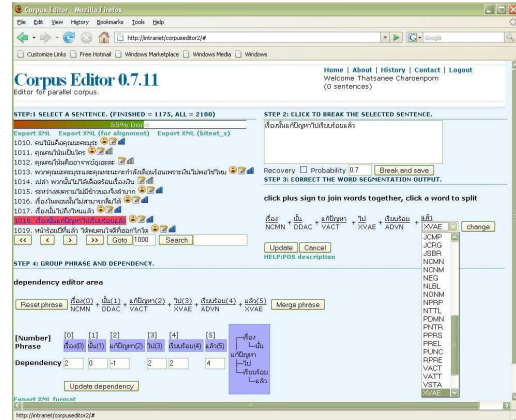
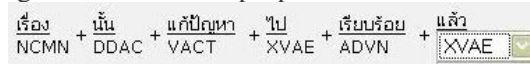


Figure 2. Syntactic Dependency Tree Annotation

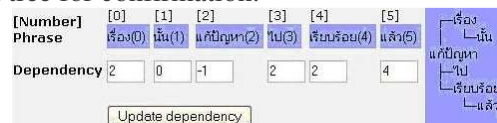
### 3.1 POS Annotation

The result from the automatic word segmentation and POS tagging program is generated with alternative POSs for revision. A dropdown list of POSs is provided for annotator to correct the POS. Since word segmentation is processed together with POS tagging, the annotator is also provided a GUI to merge or to divide the proposed word unit.



### 3.2 Syntactic Dependency Tree Annotation

The result from POS annotation in Section 3.1 is passed to define the syntactic dependency between words. The dependency is assigned to form a phrase and a sentence respectively. The final output will be marked in the XML manner and shown as a tree for confirmation.



## 4 Conclusion

Our current work on the web-based collaborative tool for Asian WordNet construction and tool for Syntactic Dependency Tree Annotation are developed as an open platform for online contribution. A user-friendly interface and self-organizing utilities are intentionally prepared to support the online collaborative work.

## References

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergit Robkop, and Hitoshi Isahara. 2007. *Collaborative Platform for Multilingual Resource Development and Intercultural Communication*, IWIC2007, Springer, LNCS4568:91-102.

# Japanese Effort Toward Sharing Text and Speech Corpora

**Shuchihi Itahashi\*+**

\*National Institute of Informatics  
2-1-1 Hitotsubashi, Chiyoda-ku,  
Tokyo, Japan 101-8430  
itabashi@nii.ac.jp

**Koiti Hasida+**

+National Institute of Advanced  
Industrial Science and Technology  
1-1-1 Umezono, Tsukuba, Ibaraki,  
Japan 305-8568  
Hasida.k@aist.go.jp

## Abstract

This report introduces the activities of the two organizations related to collection and distribution of text and speech corpora in Japan. One is the Language Resource Association (GSK) and the other is NII-Speech Resources Consortium (NII-SRC).

## 1 Introduction

Although the need for shared speech and text data has long been acknowledged, its realization has been slow to develop in Japan. The Language Resource Association (GSK) was established in 1999 in order to share and distribute text and speech corpora. It was renovated as an NPO in 2003 with an emphasis on text corpora. The National Institute of Informatics (NII) has decided to initiate the Speech Resources Consortium (SRC) in 2006, with the goal of creating future value in information media, particularly speech media.

## 2 NII-Speech Resources Consortium

The National Institute of Informatics (NII) was founded in Tokyo, Japan in April 2000 as an inter-university research institute organized to conduct comprehensive research on informatics and to develop an advanced infrastructure for disseminating scientific information. As a part of promoting the missions, NII decided in 2006 to initiate the Speech Resources Consortium (SRC)

for creating future value in information media, particularly speech media. NII is promoting this consortium together with GSK (Itahashi and Oh-suga, 2006).

### 2.1 SRC objective and activities

The SRC aims at the collection, distribution, investigation, research, and standardization of electronic data and software tools that are necessary for the development of science, education, and industry concerning speech. The consortium will contribute to the development of the information society through these activities.

The SRC investigates the existing speech resources, catalogues them, and shows them on its homepage in order to promote the research and development of speech information processing. It also requests research institutions to offer their speech resources to the SRC. Based on these activities, it urges the distribution, promotion, and publicity of speech resources. The consortium will conduct the additional production and distribution of the speech resources that are frequently requested. It also conducts investigations and research on speech resources. Users will be able to obtain the speech resources or data they need and use them through simple processes offered by the SRC. SRC distributes 23 speech corpora.

### 2.2 Organization

The SRC is made up of a chairperson, a researcher, an adviser, a secretary, and a Speech Corpora Promotion Committee. The committee works to promote the development of the SRC. Around 15 members have been invited to join the committee from the fields of speech processing,

linguistics, acoustics, speech and language corpus creation, and speech and language resource provision. The committee meets a few times a year.

### 3 Language Resource Association (GSK)

The GSK was renovated as an NPO in 2003 and is qualified as a corporate body and can mediate between the producer and users of a language corpora (Hasida and Tanaka, 2006).

#### 3.1 GSK objective and activities

The GSK aims at almost the same objectives as those of the NII-SRC mentioned in the preceding section, including both the text and speech corpora, but the text corpora are the main concern of GSK at present.

Prof. Y. Mikami of Nagaoka University of Technology proposed a project on the “Construction of Networks for Asian Linguistic Information Technology Resources” together with GSK. This proposal was adopted as a three-year project starting from fiscal 2005. It is supported by the Science and Technology Promotion & Coordination Fund, and Yen15 million (\$125,000) is available for GSK. The mission of the project is to create a network of qualified Asian partners to specify and support the development of high priority language resources for Asian languages. As a part of this project, the GSK organized the “Asian Language Resources Workshop 2007” in March. Twenty one people from 13 countries participated in the workshop.

The GSK activities are almost the same as those of the NII-SRC. The GSK is made up of a president, two vice presidents, 11 board members, 25 steering committee members, a secretary, and two office clerks. GSK supplies two text corpora and a speech corpus; it plans to distribute seven more text corpora soon.

#### 3.2 Corpora distribution system

We have three systems of corpora distribution to be conducted by the NII-SRC and GSK. (1) No-fee distribution: As a rule, the cost of handling the corpora is to fall on the user, although the corpus itself is free of charge. (2) Agency: The producers of the corpora entrust the SRC/GSK with receiving requests from the users. The SRC/GSK advertises the corpora to speech researchers

through the homepage. It mediates the user’s requests to the producer or provider of the corpora. (3) Fee-based distribution: Making speech corpora usually requires some money, including royalties. Some corpora cost users a certain amount of money to obtain, although they are not so expensive.

### 4 Present Issues

So far, we have established the GSK and NII-SRC for the text and speech corpora, respectively, while the LDC and ELRA distributes both the speech and text corpora.

The GSK is supported by a project until the end of this fiscal year, but it will lose its financial support at that point. The GSK has a relatively small amount of corpora to be distributed, so it is still too early to stand on its own feet. The NII-SRC activities will be supported by a project for a few more years, but it can not gain profit from distributing the corpora created by the researchers outside of NII. We will search for better ways by which both the NII-SRC and GSK will be able to act as corpora agents like the LDC or ELRA. There are some more organizations related to language resources such as ATR, NICT, and NIJL. We need much more collaboration and coordination among these.

### 5 Conclusion

This report explained the activities of the GSK and the NII-Speech Resources Consortium (NII-SRC). GSK and NII-SRC will facilitate the distribution, promotion, and publicity of the language resources, and in so doing, will contribute to the information society in Japan and in Asia. For further information, please refer to the following URLs.

<http://research.nii.ac.jp/src/eng/index.html>

[http://www.gsk.or.jp/index\\_e.html](http://www.gsk.or.jp/index_e.html)

### References

- S. Itahashi, T. Ohsuga. 2006. Introduction of NII-Speech resources Consortium, *Proc. Oriental COCOSA-2006*, Penang, Malaysia: 38-43.
- K. Hasida, H. Tanaka. 2006. Nonprofit Organization “Language Resource Association (GSK)”, (in Japanese) *Japanese Linguistics*, 20:107-110.

## Author Index

Bali, Kalika .....	89	L, Sobha .....	89
Bandyopadhyay, Sivaji .....	1	Lee Nagano, Robin .....	25
Baskaran, Sankaran .....	89	Li, Wenjie .....	17
Basu, Anupam .....	57	Lu, Qin .....	17
Bhattacharya, Tanmoy .....	89		
Bhattacharyya, Pushpak .....	89	Maekawa, Kikuo .....	101
		Matsuda, Makiko .....	25
Caminero, Rizza .....	49	Mikami, Yoshiki .....	25, 49
Charoenporn, Thatsanee .....	105	Mitra, Pabitra .....	9
Chimeddorj, Odbayar .....	97	Mokarat, Chumpol .....	105
Cui, Gaoying .....	17	Murthy, Kavi Narayana .....	41
Dasgupta, Tirthankar .....	57	Nishimura, Ryo .....	95
Dinh-Thi-Phuong, Thu .....	103		
Diwakar, Synny .....	57	Okada, Yoshihiro .....	95
Ekbal, Asif .....	1	Phyo, Thin Zar .....	33
		Prasad, Rashmi .....	73
G, Uma Maheswara Rao .....	85		
Goto, Hiroki .....	25	R.J., Rama Sree .....	85
		Rau, D. Victoria .....	81
Hasida, Kôiti .....	107	Riza, Hammam .....	93
Hayase, Yoshikazu .....	25		
Hoang-Anh, Vie .....	103	S, Rajendran .....	89
Htay, Hla Hla .....	41	Saha, Sujan Kumar .....	9
Husain, Samar .....	73	Sarkar, Sudeshna .....	9
Hussain, Sarmad .....	99	Sharma, Dipti .....	73
Huynh-Quyet, Thang .....	103	Shukla, Sambit .....	57
		Sornlertlamvanich, Virach .....	105
Isahara, Hitoshi .....	105		
Itahashi, Shuichi .....	107	Tai, Yin-Sheng .....	81
		Takahashi, Tomoe .....	25
Jaimai, Purev .....	97	Thayaboon, Suphanut .....	105
Jha, Girish Nath .....	89		
Joshi, Aravind .....	73	Watanabe, Yasuhiko .....	95
		Webber, Bonnie .....	65
K V, Subbarao .....	89		
K, Saravanan .....	89	Yang, Meng-Chien .....	81
K.V., Madhu Murthy .....	85		
Ko Ko, Wunna .....	33	Zeyrek, Deniz .....	65
Kumar, Sandeep .....	57		