Language Resources in the Asian Proper—A Decade of Oriental COCOSDA

Chiu-yu Tseng
Institute of Linguistics, Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw
Website: http://phslab.ling.sinica.edu.tw
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Events
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Language Resources—Backbone of Speech Technology

1. Speech technology initiated methodology innovations in the 80’s
   ○ speech corpora
   ○ data-drive, statistically based modeling.

2. Critical issues of speech corpora
   ○ Design, annotation, platform, assessment

3. Language specific vs. cross-linguistic
   ○ What and how
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Necessity of Speech Corpus

Speech Research

↑

Objectivity of Research

↑

Openness to the Public

↓

Preserving Cultural Legacy

↓

Related Information

→

Speech Data

↓

Preservation of Spoken Language
1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Organizing Creation & Utilization of Speech Corpora

Creation of speech corpora needs some cost. Utilization needs a system to distribute corpora. Some activities started early in 1990s.

1991  COCOSDA
1992  LDC in U.S.A.
1995  ELRA in Europe
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
COCOSDA International Committee for Coordination and Standardization of Speech Databases

- Established in Europe in 1991

- Purpose
  - To promote the development of spoken language corpora for building and/or evaluating spoken language technology
  - To offer coordination of projects and research efforts to improve their efficiency.
COCOSDA Organization and Members

1. convener
   - Dr./Professor Dafydd Gibbon, Bielefeld University, Germany (2005--)

2. 11 areas
   - England, France, Germany, Holland, Italy, Japan, South Africa, Taiwan, U.S.A, China, Greece

3. Activities
   - Satellite meeting at major conferences, e.g., Eurospeech, Interspeech, REC..etc.
Back in 1993, in Asia……..

- Professor Hiroya Fujisaki, Tokyo University, Japan

1. Population 3.8 billion (http://en.wikipedia.org/wiki/)

2. Language varieties

   Altaic (Mongolia, Korea, Japan…)

   Sino-Tibetan (China, Cambodia, Laos, Thailand, Vietnam…)

   Austronesian (Taiwan, Philippines, Indonesia, Malaysia, and Oceanic islands)

   Indian (Hindi, Bangali, Urdu…..)
Major Language Families in Asia Spoken by Population 3.8 billion
(http://en.wikipedia.org/wiki/)
Language Families of Asian Languages (Ethnologue.com)

1. Austronesian (1268 languages): Malay, Indonesian, etc.
2. Sino-Tibetan (403): Chinese, Tibetan, Burmese, etc.
4. Tai-Kadai (76): Thai, Lao, etc.
5. Dravidian (73): Tamil, Telugu, etc.
6. Altaic (66): Mongolian, Turkic, Korean, etc.
8. *Indo-European* (449)
Asian Countries & Territories

- East Asia, South-East Asia, Indo-Tibetan area, Arabic area, NIS area (43+1) Country (as of 1990-1993)
- Area ($10^3$km$^2$)
- Population (million)
- Density (/km$^2$)
- Major Languages
### East Asia

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (10^3 km²)</th>
<th>Population (million)</th>
<th>Density (/km²)</th>
<th>Major Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>9,597</td>
<td>1,155.80</td>
<td>120</td>
<td>Chinese</td>
</tr>
<tr>
<td>DPR of Korea</td>
<td>121</td>
<td>22.20</td>
<td>183</td>
<td>Korean</td>
</tr>
<tr>
<td>Japan</td>
<td>378</td>
<td>123.92</td>
<td>328</td>
<td>Japanese</td>
</tr>
<tr>
<td>Mongolia</td>
<td>1,567</td>
<td>2.25</td>
<td>1</td>
<td>Mongolian</td>
</tr>
<tr>
<td>Republic of Korea</td>
<td>99</td>
<td>43.27</td>
<td>437</td>
<td>Korean</td>
</tr>
<tr>
<td>Taiwan</td>
<td>36</td>
<td>2.68</td>
<td>74</td>
<td>Chinese</td>
</tr>
</tbody>
</table>
## South-East Asia

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
<th>GDP per Capita</th>
<th>Dual Language Speakers</th>
<th>Language(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brunei</td>
<td>6</td>
<td>0.27</td>
<td>45</td>
<td>Malay, English</td>
</tr>
<tr>
<td>Cambodia</td>
<td>181</td>
<td>8.44</td>
<td>47</td>
<td>Cambodian</td>
</tr>
<tr>
<td>Indonesia</td>
<td>1,905</td>
<td>187.77</td>
<td>99</td>
<td>Indonesian</td>
</tr>
<tr>
<td>Laos</td>
<td>237</td>
<td>4.26</td>
<td>18</td>
<td>Laotian</td>
</tr>
<tr>
<td>Malaysia</td>
<td>330</td>
<td>18.33</td>
<td>56</td>
<td>Malay</td>
</tr>
<tr>
<td>Maldives</td>
<td>0.3</td>
<td>0.22</td>
<td>733</td>
<td>Divehi</td>
</tr>
<tr>
<td>Myanmar</td>
<td>677</td>
<td>42.56</td>
<td>63</td>
<td>Burmese</td>
</tr>
<tr>
<td>Singapore</td>
<td>0.62</td>
<td>2.76</td>
<td>4,450</td>
<td>Malay, Eng. Chin. Tamil</td>
</tr>
<tr>
<td>Thailand</td>
<td>513</td>
<td>56.92</td>
<td>111</td>
<td>Thai</td>
</tr>
<tr>
<td>The Philippines</td>
<td>300</td>
<td>62.87</td>
<td>210</td>
<td>Pilipino, English</td>
</tr>
<tr>
<td>Viet Nam</td>
<td>332</td>
<td>68.18</td>
<td>205</td>
<td>Vietnamese</td>
</tr>
</tbody>
</table>
## Indo-Tibetan Area

<table>
<thead>
<tr>
<th>Country</th>
<th>Area ($10^3$km$^2$)</th>
<th>Population (million)</th>
<th>Density (/km$^2$)</th>
<th>Major Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>652</td>
<td>16.43</td>
<td>25</td>
<td>Bashto, Dari</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>144</td>
<td>118.75</td>
<td>825</td>
<td>Bengali</td>
</tr>
<tr>
<td>Bhutan</td>
<td>47</td>
<td>1.55</td>
<td>33</td>
<td>Dzongkha</td>
</tr>
<tr>
<td>India</td>
<td>3,288</td>
<td>849.64</td>
<td>258</td>
<td>Hindi+13, English</td>
</tr>
<tr>
<td>Nepal</td>
<td>141</td>
<td>19.61</td>
<td>139</td>
<td>Nepalese</td>
</tr>
<tr>
<td>Pakistan</td>
<td>796</td>
<td>115.52</td>
<td>145</td>
<td>Urdu, English</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>66</td>
<td>17.24</td>
<td>261</td>
<td>Singhalese, Eng.</td>
</tr>
</tbody>
</table>
## Arabic Area

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
<th>Percentage</th>
<th>Residents</th>
<th>Language(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain</td>
<td>765</td>
<td>0.68</td>
<td>0.52</td>
<td>Arabic</td>
</tr>
<tr>
<td>Cyprus</td>
<td>79</td>
<td>0.71</td>
<td>0.52</td>
<td>Greek, Turkish, English</td>
</tr>
<tr>
<td>Iran</td>
<td>35</td>
<td>57.73</td>
<td>1,648</td>
<td>Persian</td>
</tr>
<tr>
<td>Iraq</td>
<td>45</td>
<td>19.58</td>
<td>438</td>
<td>Arabic</td>
</tr>
<tr>
<td>Israel</td>
<td>237</td>
<td>4.98</td>
<td>21</td>
<td>Hebrew</td>
</tr>
<tr>
<td>Jordan</td>
<td>42</td>
<td>4.15</td>
<td>98</td>
<td>Arabic</td>
</tr>
<tr>
<td>Kuwait</td>
<td>117</td>
<td>2.10</td>
<td>18</td>
<td>Arabic</td>
</tr>
<tr>
<td>Lebanon</td>
<td>275</td>
<td>2.75</td>
<td>10</td>
<td>Arabic</td>
</tr>
<tr>
<td>Oman</td>
<td>7</td>
<td>1.56</td>
<td>212</td>
<td>Arabic</td>
</tr>
<tr>
<td>Qatar</td>
<td>35</td>
<td>0.38</td>
<td>11</td>
<td>Arabic</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>8</td>
<td>16.93</td>
<td>2,150</td>
<td>Arabic</td>
</tr>
<tr>
<td>Syria</td>
<td>70</td>
<td>12.99</td>
<td>185</td>
<td>Arabic</td>
</tr>
<tr>
<td>United Arab Emeritus</td>
<td>19</td>
<td>1.63</td>
<td>84</td>
<td>Arabic</td>
</tr>
<tr>
<td>Yemen</td>
<td>21</td>
<td>11.28</td>
<td>528</td>
<td>Arabic</td>
</tr>
<tr>
<td>(Turkey)</td>
<td>74</td>
<td>57.33</td>
<td>779</td>
<td>Turkish</td>
</tr>
<tr>
<td>Country</td>
<td>Area ($10^3$km²)</td>
<td>Population (million)</td>
<td>Density (/km²)</td>
<td>Major Languages</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>2,717</td>
<td>16.79</td>
<td>6</td>
<td>Kazakh, Russian</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>199</td>
<td>4.45</td>
<td>22</td>
<td>Kyrgys</td>
</tr>
<tr>
<td>Tadzhikistan</td>
<td>143</td>
<td>5.47</td>
<td>37</td>
<td>Tadzhik</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>488</td>
<td>3.71</td>
<td>8</td>
<td>Turkmen</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>447</td>
<td>20.71</td>
<td>46</td>
<td>Uzbek</td>
</tr>
</tbody>
</table>
Features of Asian Languages

1. Language and Language families—Many varieties

2. Orthographic systems--Large varieties of letters/characters

3. Tonal vs. Non-tonal languages—Both type exist.

4. Spacing between words—Only some languages keep space.

5. Romanization--Non-unique systems are used.

6. Many languages still do not have working writing system. Speech is the only linguistic mode.
Letters, Tone & Word Order

- Proper letters
  - Burmese, Chinese, Japanese, Khmer, Korean, Thai, etc.

- Latin letters
  - Indonesian, Malay, Vietnamese, etc.

- Tonal languages
  - Burmese, Chinese, Lao, Thai, Vietnamese, etc.
  - Speaking population 1,552 million vs. African tone language population 840 million

- Word order
  - SOV, SVO, VSO, VOS
Word Boundary by Spacing in Text

- No space between words
  - Burmese, Chinese, Japanese, Khmer, Lao, Thai, etc.

- Space between words
  - Indonesian, Malay, Mongolian, Vietnamese, etc.
Necessity of Oriental COCOSDA

- Asia is a multilingual region.
- Diversity of the languages is larger than Europe.
- Speech researches were emerging.
- Speech corpora were required.
- Cooperation among countries was necessary.
- Organizations for speech corpora were needed.
Strategies of Oriental COCOSDA

1. Founding a forum of speech corpora

2. Establishing regional consortia:
   - GSK,
   - SITEC,
   - Chinese LDC, CCC,
   - NII-SRC

3. Promoting collaboration among the consortia
Oriental Chapter of COCOSDA

Mission

- To promote speech research on oriental languages via
  - 1. exchange ideas
  - 2. share information
  - 3. discuss regional matters on SLP related issues
    - Creation
    - Utilization
    - Dissemination of spoken language corpora of oriental languages
  - 4. assessment methods of speech input/output systems
History of Oriental COCOSDA

- Proposed in 1994
- Preparatory meeting in Hong Kong in 1997
- Annual workshops held since 1998.
Goals of Oriental COCOSDA

1. Initiating Speech Resources Consortium in each country.

2. Establishing Asian Network among the Consortia.

3. Creating multilingual corpus of semantically similar contents.
### Organization

- **Convener:**
  - Dr. Chiu-yu TSENG (2006-)
  - Dr. Shuichi ITAHASHI (1998-2005)

- **Advisory members:**
  - 1 from China, Japan and Korea each

- **28 committee members from 14 region/country (2007):**
  - China, Hong Kong, India, **Indonesia**, Japan, Korea, Malaysia, Mongolia, Nepal, Pakistan, Singapore, Taiwan, Thailand and Vietnam.
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Events
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Major Functions and Flagship Events of Oriental-COCOSDA

- **Major function**
  - Networking among members
  - Attracting new members

- **Flagship events**
  - Annual international workshops since 1998
Oriental COCOSDA Venues and Organizers

- T.F. Zheng (China)
- S.S. Agrawal (India)
- Thanaruk T. (Thailand)
- A. Pandian (Malaysia)
- K.T. Lua (Singapore)
- S. Itahashi (Japan)
- Y-J Lee (Korea)
- L.S. Lee (Taiwan)
- C.K. Chan (Hong Kong)
- C.M. Luong (Vietnam)
- H. Riza (Indonesia)

- Around 30 papers with 60 participants
Oriental COCOSDA WORKSHOP—
International Workshop on East-Asian Language Resources, Evaluation and Assessment

1998  1st Meeting, Tsukuba, Japan (30 papers, 54 participants)
1999  2nd Meeting, Taipei, Taiwan (44, 120)
2000  3rd Meeting, Beijing, China (8, 20)
2001  4th Meeting, Taejon, Korea (11, 25)
2002  5th Meeting, Hua Hin, Thailand (24, 96) + SNLP
2003  6th Meeting, Sentosa, Singapore (28, 60 ) + PACLIC
2004  7th Meeting, Delhi, India (55, 150) + iSTEPS
2005  8th Meeting, Jakarta, Indonesia (24, 65)
2006  9th Meeting, Penang, Malaysia (34, 60)
2007  10th Meeting, Hanoi, Vietnam (34, 75)
Future Oriental-COCOSDA Conferences

Oriental COCOSDA 2008
25-27 Nov. 2008
ATR Spoken Language Translation Res. Labs.
Kyoto, Japan

http://www.slc.atr.jp/o-cocosda/

Abstract submission: Aug. 29
Notification of acceptance: Sep. 19
Final manuscript: Oct. 24
Future Oriental-COCOSDA Conferences

● Oriental-COCOSDA 2009
  ○ Late August at Urumuqi, Xinjiang Autonomous Region of China, back-to-back with NCMMSC (National Conference on Man-Machine Speech Communication, China)

● Oriental-COCOSDA 2010
  ○ Katmandu, Nepal
1. Currently the flagship organization of COCOSDA
   - Euro-COCOSDA
   - Oriental-COCOSDA
   - African-COCOSDA (under preparation)

2. Differ from Write, WordNet…etc
   - Region oriented and uniquely Asian
   - Speech database focused vs. text corpora
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Other Asian Activities since Oriental-COCOSDA

1997 Oriental COCOSDA
1999 GSK (Language Resource Association) in Japan
2001 SITEC in Korea
   (Speech Information Technology & Industry Promotion Center)
2002 Chinese LDC
   CCC (Chinese Corpus Consortium) in China
2006 NII-SRC in Japan
   (National Institute of Informatics, Speech Resources Consortium)
Japanese Activities

**GSK**: Language Resource Association
- Launched in 1999
- Renovated as an NPO in 2003
- Project accepted in 2005 for 3 years
- Emphasizing written text corpora

**NII-SRC**: launched in 2006 for speech corpora
Standardization in Japan

1. **Open Software Tools**: Julius, Galatea, etc.


JEITA: Japan Electronics and Information Technology Industries Association

JEIDA: Japan Electronic Industry Development Association
Korea

SITEC (Speech Information Technology & Industry Promotion Center)
Founded in 2001 (Korean LDC/ELRA)
Wonkwang University as host organization
(7 full-time staffs)
Chinese LDC

Launched in 2002
Creation of linguistic corpora
Management & distribution of language resources
Promotion of sharing language resources

*Chinese Corpus Consortium (CCC)*
Future of Oriental COCOSDA

1. Long-Term collaboration among regional activities

2. Cooperative creation of speech corpora and planning of resource sharing

3. More promotion of speech research in Asia
   - Future conference sites:
     Xinjiang Uygur Autonomous Region of China, Nepal, Mongolia, etc.
Time for Resource Sharing

- A-STAR
- AESOP

- Many more to be organized w.r.t. what to be included and how to construct cross-linguistic platforms
Future Prospects: Global Speech Corpus

Digits, digit strings, days of the week, months, time, salutations, yes/no, well-known proper nouns (person names, cities, companies), well-known stories, phonetically-balanced sentences, etc. common to all languages.
Utterance Content

Items widely understood in the world:
10 Digits, 12 Months of the year,
7 Days of the week, 4 Words on Weather,
6 Phrases of Greetings, 3 Words of Replies,
4 Words on time.

“North Wind” from Aesop’s Fables
Features of the proposed corpus

Containing various Asian Languages

With the same semantic content

Recorded in a sound-proof room
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Visibility of O-C in Neighboring Communities

- Holding O-C conferences with other local conferences
  - 2000  3rd O-C Workshop, Beijing, China (8, 20) as satellite of ISCSLP2000
  - 2002  5th O-C Workshop, Hua Hin, Thailand (24, 96) with SNLP
  - 2003  6th O-C Workshop, Sentosa, Singapore (28, 60) with PACLIC
  - 2004  7th O-C Workshop, Delhi, India (55, 150) + iSTEPS
Visibility of O-C in Neighboring Communities (2/4)

2. Other workshops held because of O-C conference—NII Symposium

- International Symposium on Asian Language Resources 2008
Visibility of O-C in Neighboring Communities (3/4)—
2nd Int’l Workshop on Language and Speech Science, (Sept.4-5, 2008 Waseda U, Tokyo, Japan

- Sept 4
  - Phonetics in L2 English
  - Spoken Language Processing
  - Automatic Assessment of L2 English
  - L2 (French, German, Japanese)

- Sept 5
  - Japanese Timing and L2
  - English Timing and L2
  - L2 (Chinese)
  - L2-Education

Participating countries

  Japan, USA, Indonesia, Kong Kong, Switzerland, Thailand
Visibility of O-C in Neighboring Communities
(4/4) -- NII Symposium, Tokyo (Nov. 29, 2008)

1. Speech and language resources of some Indian languages
   - Dr. S. S. Agrawal (KlIT College of Engineering, India),
2. Recent activities of Chinese Corpus Consortium (CCC)
   - Prof. T. Fang Zheng (Tsinghua University, China),
3. Recent activities of Speech Information Technology Promotion Center (SITEC) in Korea,
   - Prof. Y-J Lee (Wonkwang University/SITEC, Korea),
4. Speech and language resources of Indonesian languages
   - Dr. Hammam Riza (Agency for Assessment and Application of Technology (BPPT), Indonesia),
5. Some recent developments of speech resources for Vietnamese
   - Dr. Vu Tat Thang (Vietnamese Academy of Science and Technology (VAST), Vietnam),
6. Speech and language resources at NICT/ATR
   - Dr. S. Nakamura (National Institute of Communications Technology/ATR-SLC, Japan)
Closing address
Prof. S. Itahashi (NII/AIST, Japan)
O-C as One Community to Other International Activities

1. Organizing special session in neighboring conference ISCSLP2006, Singapore after the 9th O-C, Malaysia
   - Session: Multilingual Corpus Development
   - Participated by: Japan, Indian, Vietnam, Mongolia

2. Reporting O-C activities at LREC
   - LREC 2006, Italy
   - LREC 2008, Morocco
Outline

1. Necessity of Speech Corpora
2. Organizing the Creation and Utilization of Speech Corpora
3. Why COCOSDA and Oriental-COCOSDA?
4. Oriental-COCOSDA Functions and Flagship Event
5. Other Asian Activities since O-C
6. Visibility of O-C
7. Commemoration of 10 years of O-C
Oriental COCOSDA Book Project
In Commemoration of the First Decade of Sustained Activities in Asia

*Resources and Standards of Spoken Language Systems – Advances in Oriental Spoken Language Processing* – Edited by

- Shuichi ITAHASHI
  - National Institute of Informatics (NII), Japan
- Chiuyu TSENG
  - Institute of Linguistics, Academia Sinica, Taiwan
Contents of the 500-page Book:

1. Introduction
2. Outline of Oriental languages
3. Data centers and corpora
4. Speech corpora of Oriental languages
5. Performance evaluation of synthesizers and recognizers
6. Annotation and labeling
7. Software tools
8. Orthographic transcription and Romanization
9. Conclusion
   Appendix: History of Oriental COCOSDA
Oriental COCOSDA

- The Future of Asian SLP and Speech Technology Development Depends on All of Us.

- Your Sustained Participation and Contribution Is Our Lifeline.

Thank you!