Asian Activities on Speech Corpora

Chiu-yu Tseng * Shuichi ITAHASHI[!+]

* Academia Sinica, Taipei, Taiwan

[!] National Institute of Informatics (NII), Tokyo, Japan
[+]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

East Asia (6)

Country	Area ($10^3 km^2$)	Population (million)	Density ($/km^2$)	Major Languages
China	9,597	1,155.80	120	Chinese
DPR of Korea	121	22.20	183	Korean
Japan	378	123.92	328	Japanese
Mongolia	1,567	2.25	1	Mongolian
Republic of Korea	99	43.27	437	Korean
Taiwan	36	2.68	74	Chinese

South-East Asia (11)

Brunei	6	0.27	45	Malay, English
Cambodia	181	8.44	47	Cambodian
Indonesia	1,905	187.77	99	Indonesian
Laos	237	4.26	18	Laotian
Malaysia	330	18.33	56	Malay
Maldives	0.3	0.22	733	Divehi
Myanmar	677	42.56	63	Burmese
Singapore	0.62	2.76	4,450	Malay,Eng. Chin.Tamil
Thailand	513	56.92	111	Thai
The Philippines	300	62.87	210	Pilipino, English
Viet Nam	332	68.18	205	Vietnamese

Indo-Tibetan Area (7)

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

Why Speech Corpora?
Typological Features of Asian Languages (1/2)

1. Many **language families, including both tonal and non-tonal languages.**
   - Austronesian (1268 languages): Malay, Indonesian, etc.
   - Sino-Tibetan (403): Chinese, Tibetan, Burmese, etc.
   - Austro-Asiatic (169): Khmer, Vietnamese, etc.
   - Tai-Kadai (76): Thai, Lao, etc.
   - Dravidian (73): Tamil, Telugu, etc.
   - Altaic (66): Mongolian, Turkic, Korean, etc.
   - Japanese (12): Japanese, Ryukyuan, etc.

   cf. Indo-European (449) by Ethnologue.com

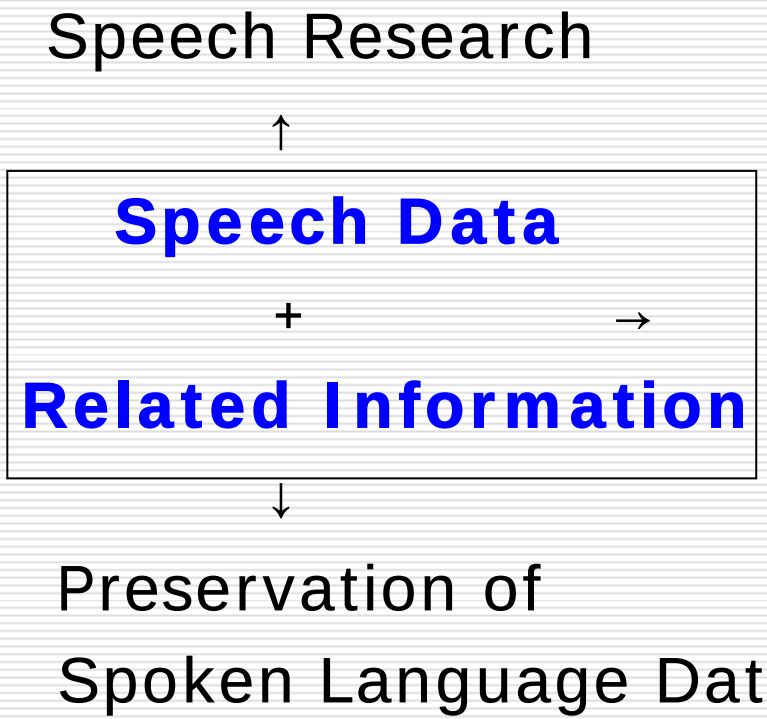
Why Speech Corpora?
Typological Features of Asian Languages (1/2)

2. Large variety of **writing** systems and formats
    1. Letters, Tone & Word Order
        1. Proper letters: Burmese, Chinese, Japanese Khmer, Korean, Thai, etc.
        2. Latin letters: Indonesian, Malaysian, Vietnamese, etc.
    2. space/no space are used
        Space between words: Indonesian, Malay, Mongolian, Vietnamese, etc.
        No space between words: Burmese, Chinese, Japanese, Khmer, Lao, Thai, etc
    3. Word order: SOV, SVO, VSO, VOS

3. Many languages do not have writing systems

4. Non-unique **Romanization** systems

Regional Necessity of Speech Corpus



Objectivity of Research

↑

Openness to the Public

↓

Preserving Cultural Legacy

Oriental COCOSDA (1/2)

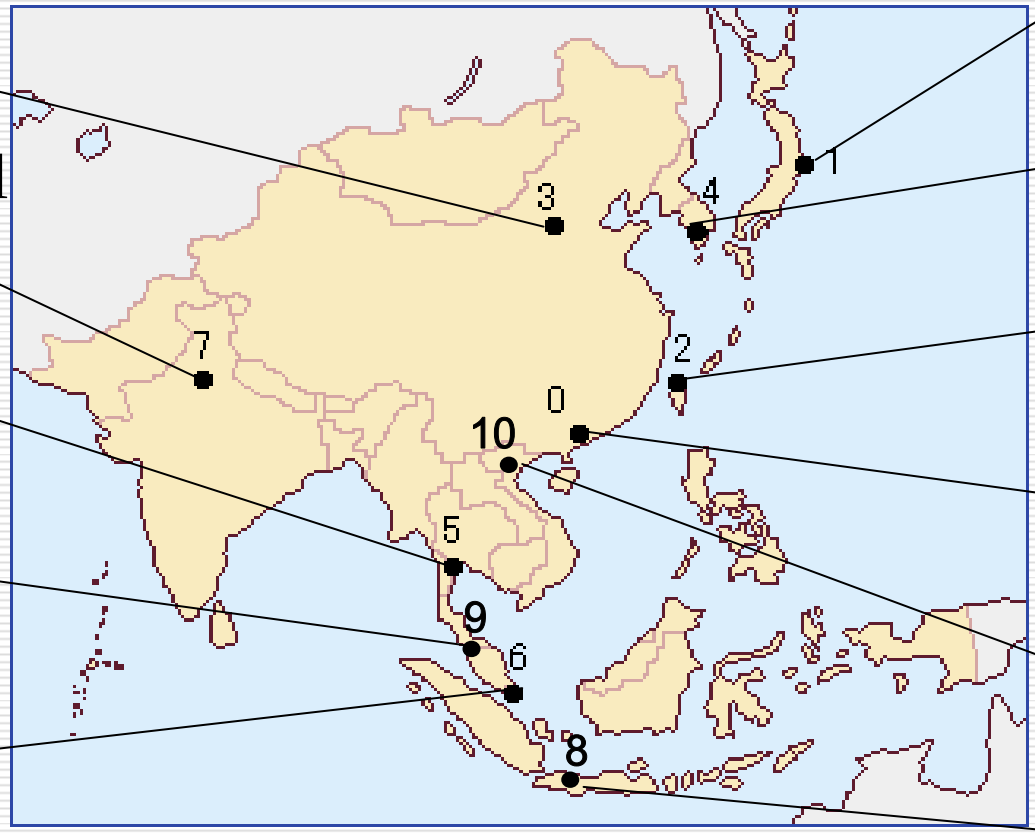
1. History

1. Proposed in 1994, to exchange ideas, share information, discuss regional issues on SLP.
2. Preparatory meeting in Hong Kong in 1997.

2. International workshops held

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)

Oriental COCOSDA 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)

Oriental COCOSDA (2/2)

3. Goals

- Initiating Speech Resources Consortium in each country.
- Establishing Asian Network among the Consortia.
- Creating multilingual corpus of semantically similar contents.

4. Mission

1. To exchange ideas, share information, discuss regional matters on creation, utilization, dissemination of spoken language corpora of oriental languages, assessment methods of speech input/output systems.
2. To promote speech research on oriental languages.

5. Strategies

1. Foundation of Oriental COCOSDA →Forum of speech corpora
2. Establishment of Regional Consortia:

   GSK, SITEC, Chinese LDC, CCC, NII-SRC
3. Collaboration among the consortia.

Asian Activities since O-C 1997

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)

Oriental COCOSDA Organization

Convenor: Chiu-yu TSENG (2006-)

S. ITAHASHI (1998-2005)

Advisory members:

Three from China, Japan, Korea

Committee members: 26 from 13 regions including

China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, Mongolia, Nepal, Singapore, Taiwan, Thailand, Vietnam.

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

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)

Oriental COCOSDA Book Project

Tentative Title:

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

Editors: S. Itahashi and C-Y Tseng

Contents

Future of Oriental COCOSDA

1. Collaboration among regional activities
    1. A-STAR
    2. AESOP
2. Cooperative creation of speech corpora
3. Promotion of speech research in Asia
4. Forming possible common platform

Upcoming Oriental-COCOSDA Workshops and other workshop

1. **Oriental-COCOSDA 2008**
   25-27 Nov. 2008
   ATR Spoken Language Translation Res. Labs.
   Kyoto, Japan
   Abstract submission: Aug. 29
   Notification of acceptance: Sep. 19
   Final manuscript: Oct. 24
   http://www/slc.atar.jp/o-cocosda
2. **International Symposium on Asian Language Resources 2008**
   28 Nov. 2008
   National Institute of Informatics (NII), Tokyo, Japan
   http://www.slc.atr.jp/o-cocosda/
   /

3. **Oriental-COCOSDA 2009**
   Xinjang University, Uygur Autonomous Region of China

Conclusion

1. Importance of speech corpora for promoting speech research.

2. Role of organizations for speech corpus creation and distribution.

3. GSK, SRC/SITEC/Chinese LDC, CCC are expected to further speech corpus creation and distribution together with Oriental COCOSDA in East Asia.

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

4. The reported book project will serve as documentation of continued Oriental-COCOSDA activities.