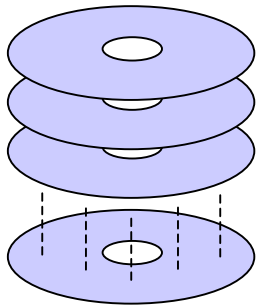
Oriental COCOSDA: Past, Present and Future

Shuichi ITAHASHI
National Institute of Informatics (NII), Tokyo, Japan
AIST, Tsukuba, Japan
Chiu-yu TSENG
Academia Sinica, Taipei, Taiwan
Satoshi NAKAMURA
ATR Spoken Language Communication Res. Labs., Kyoto, Japan

Contents

1. Necessity of Speech Corpora
2. Organizations for Speech Corpora
3. Asian Languages
4. Brief History
5. Goals & Strategies
6. Regional Activities
7. Conclusion



Necessity of Speech Corpus

Speech Research

↑

Speech Data

\+ →

Related Information

↓

Preservation of

Spoken Language Data

Objectivity of Research

↑

Openness to the Public

↓

Preserving Cultural Legacy

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

COCOSDA

International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment

Workshops held annually at Interspeech

Cocosda promotes the development of spoken language corpora for building and/or evaluating spoken language technology and offers coordination of projects and research efforts to improve their efficiency.

Features of Asian Languages

1. Many languages belong to different **language families.**
2. Variety of **orthographic** systems

   Various letters/characters used
3. Some **tonal** languages
4. **No space** between words in some languages
5. Non-unique **romanization** systems

Language Families of Asian Languages

1. Austronesian (1268 languages): Malay, Indonesian, etc.
2. Sino-Tibetan (403): Chinese, Tibetan, Burmese, etc.
3. Austro-Asiatic (169): Khmer, Vietnamese, etc.
4. Tai-Kadai (76): Thai, Lao, etc.
5. Dravidian (73): Tamil, Telugu, etc.
6. Altaic (66): Mongolian, Turkic, Korean, etc.
7. Japanese (12): Japanese, Ryukyuan, etc.

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

Letters, Tone & Word Order

1. Proper letters: Burmese, Chinese, Japanese, Khmer, Korean, Thai, etc.
2. Latin letters: Indonesian, Malay, Vietnamese, etc.
3. **Tonal** languages: Burmese, Chinese, Lao, Thai, Vietnamese, etc.
4. Word order: SOV, SVO, VSO, VOS

Word boundary in text

1. No space between words: Burmese, Chinese, Japanese, Khmer, Lao, Thai, etc.
2. Space between words: Indonesian, Malay, Mongolian, Vietnamese, etc.

Asian Activities

1994, 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

Proposed in 1994, to exchange ideas, share information, discuss regional issues on SLP.

Preparatory meeting in Hong Kong in 1997.

Annual workshops held since 1998 in Japan, Taiwan, China, Korea, Thailand, Singapore, India, Indonesia.

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.

Oriental COCOSDA Mission

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, and

To promote speech research on oriental languages.

Goals of Oriental COCOSDA

1. Initiating Speech Resources Consortium in each country.
2. Establishment of Asian Network among the Consortia.
3. Creation of multilingual corpus of semantically similar contents.

Strategies of Oriental COCOSDA

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

Oriental COCOSDA Organization

Convenor: Chiu-yu TSENG (2006-)

S. ITAHASHI (1998-2005)

Advisory members:

Three from China, Japan, Korea

Committee members: 21 from 10 regions including

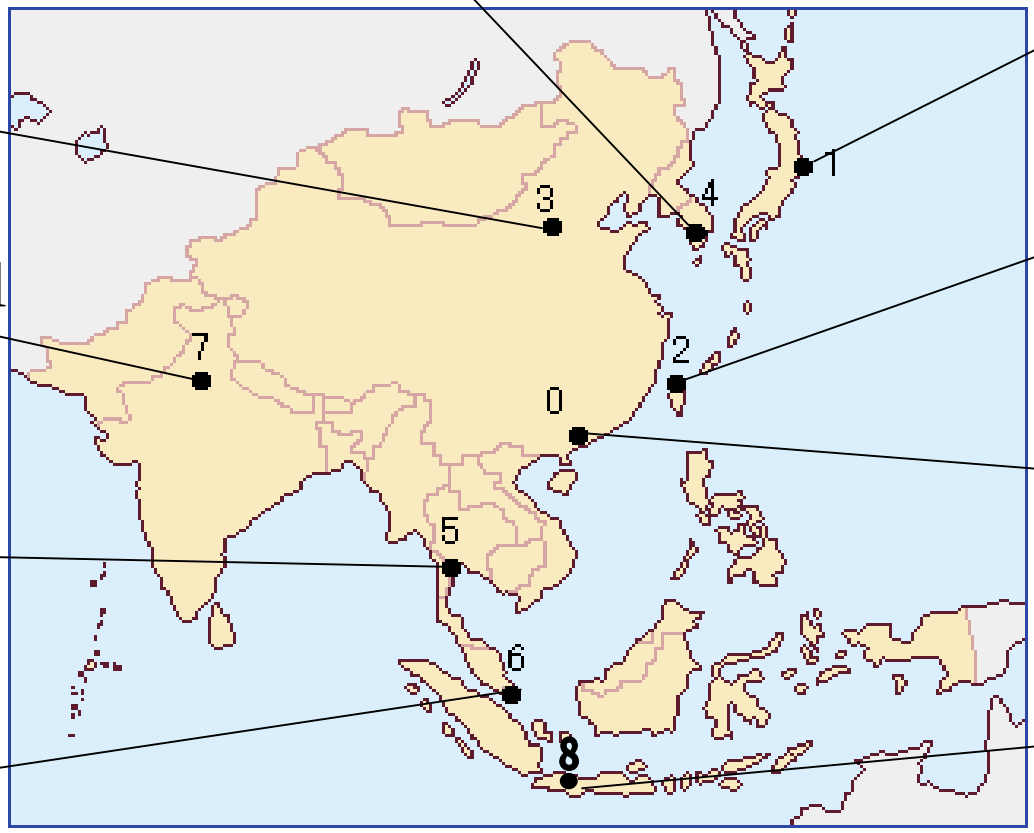
China, Hong Kong, India, Indonesia, Japan, Korea, Mongolia, Singapore, Taiwan, Thailand.

International Workshop on East-Asian Language Resources and Evaluation

- Oriental COCOSDA WORKSHOP -

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, iSTRANS
2005 8th Meeting, Jakarta, Indonesia (24, 65)

Oriental COCOSDA Organizers

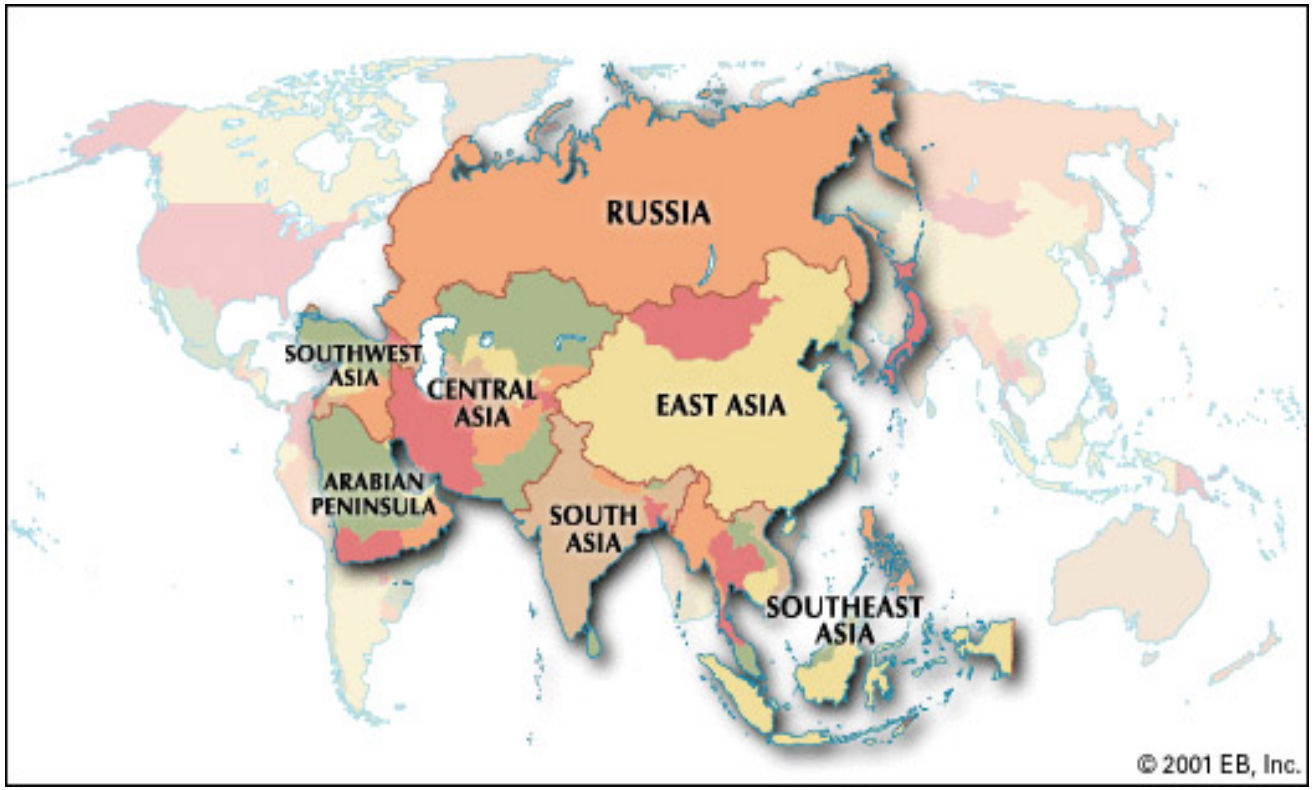


Participation

0. China, Japan, Korea, Taiwan (CJKTw), Hong Kong (HK)
1. CJKTw
2. CJKTw, Thailand (Th), France (F), U.S.A.
3. CJKTw, Th, Mongolia (Mg)
4. CJKTw, Th, Australia (Au)
5. CJKTw, Th, India (Id), Indonesia (Is), Guam
6. CJKTw, Th, Id, Is, Singapore (S)
7. CJKTw, Id, Is, S, Au, F, U.S.A.
8. CJKTw, Th, Is, Malaysia, Mg, HK

Some Regional Activities

Japan
Korea
China
Hong Kong
Mongolia
Singapore
Taiwan
Thailand
India
Indonesia



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.

2) Standard of Speech Synthesis System Performance Evaluation Methods by JEITA (2003)

3) Standard of Symbols for Japanese Text-To-Speech Synthesizer by JEIDA (2000)

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 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

Future of Oriental COCOSDA

1. Collaboration among regional activities
2. Cooperative creation of speech corpora
3. Promotion of speech research in Asia

Future conference sites:

Malaysia, Vietnam, Mongolia,

Xinjang 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
4. 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/

Oriental COCOSDA 2006

9-11 Dec. 2006

Universiti Sains Malaysia

Penang, Malaysia

Abstract submission: Aug. 5

Notification of acceptance: Aug. 26

Final manuscript: Sep. 30

http://www.usm.my/o-cocosda/

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 (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.

Arabic Area (14+1)

Bahrain	0.68	0.52	765	Arabic
Cyprus	9	0.71	79	Greek,Turkish, English
Iran	1,648	57.73	35	Persian
Iraq	438	19.58	45	Arabic
Israel	21	4.98	237	Hebrew
Jordan	98	4.15	42	Arabic
Kuwait	18	2.10	117	Arabic
Lebanon	10	2.75	275	Arabic
Oman	212	1.56	7	Arabic
Qatar	11	0.38	35	Arabic
Saudi Arabia	2,150	16.93	8	Arabic
Syria	185	12.99	70	Arabic
United Arab Emeritus	84	1.63	19	Arabic
Yemen	528	11.28	21	Arabic
(Turkey	779	57.33	74	Turkish)

NIS Area (5)

Country	Area ($10^3 km^2$)	Population (million)	Density ($/km^2$)	Major Languages
Kazakhstan	2,717	(16.79)	6	Kazakh, Russian
Kyrgystan	199	4.45	22	Kyrgys
Tadzhikistan	143	5.47	37	Tadzhik
Turkmenistan	488	(3.71)	8	Turkmen
Uzbekistan	447	(20.71)	46	Uzbek