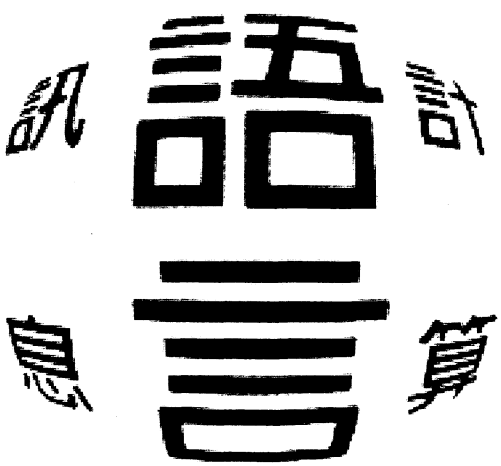
PACLIC 17

Language, Information and Computation



1-3 October, 2003
Sentosa, Singapore

Edited by
Dong Hong Ji and Kim Teng Lua

COLIPS PUBLICATIONS

Language, Information and Computation

Proceedings of 17th Pacific Asia Conference

1-3 October, 2003, Sentosa, Singapore

Edited by
Dong Hong Ji and Kim Teng Lua

COLIPS PUBLICATIONS
ISBN

©2003 Chinese and Oriental Languages Information Processing Society

All right reserved. No part of this publication may be
reproduced, stored in a retrieval system, or
transmitted, in any form or by any means, electronic,
mechanical, photocopying, recording or otherwise,
without the prior permission of the publisher.

Published by COLIPS PUBLICATIONS
c/o School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543

Welcome Message from Local Chair, Dr Lua Kim Teng

Dear Friends and Colleagues,

A very warm welcome to the beautiful Sentosa Resort Island of Singapore!

The O-COCOSDA2003 and PACLIC17 are major events in the Asian natural language research arena. This is the seventh in the series of O-COCOSDA or Oriental International Coordinating Committee on Speech Databases and Speech I/O System Assessment and the seventeenth PACLIC or Pacific Asia Conference on Language, Information and Computation. We are very honor that Singapore is chosen to be the venue to host both meetings and colips is the organizer.

COCOSDA is an international workshop held annually by the oriental chapter of The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques for Speech Input/Output. The first meeting was held in Hong Kong and then the past five conferences were held in Japan, Taiwan, China Mainland, Korea and Thailand. After the Singapore meeting, I understand that the next meeting, O-COCOSDA2004 will be held in New Delhi, India. The meeting will be organized under the leadership of Dr Shyam S. Agrawal, Emeritus Scientist, Central Scientific Instruments, Delhi. I wish that every one of us will support the meeting and make it a success!

The PACLIC series of conference is an annual meeting of scholars in theoretical and computational linguistics from the Pacific Asia region. PACLIC aims to cover all aspects of both theoretical and computational linguistics, including morphology, phonology, syntax, semantics, pragmatics, discourse analysis, typology, corpus linguistics, formal grammar theory, natural language processing, natural language systems and related computer applications.

We find that by holding the two conferences at the same venue side-by-side will allow direct interaction between members of the two communities, the speech and the text and NLP community. Perhaps, this synergy will create sparks that will be most beneficial to both research areas.

We received more than 100 submissions from research in Asia and many other parts of the words for the current joint meetings of O-COCOSDA2003 and PACLIC17. Only about 50% of the papers are selected for oral presentations. From the papers presented in the two volumes of proceedings, you will find that we are working hard to maintain high academic standards for the conferences.

I wish to thank the steering committee members: Dr Shyam Agrawal, Nick Campbell, Qiang Huo, Dawa Idomuso, Shuichi Itahashi, Aijun Li, Lin Shan Lee, Yongju Lee, Yung-Hwan Oh, Virach Sornlertlamyanich, Hsiao-Chuan Wang, Jialu Zhang (above for O-COCOSDA2003), Huang Chu-Ren, Akira Ikeya, Ik-Hwan Lee, Byung-Soo Park, Benjamin T'sou (above are co-chairs for PACLIC17) for their assistance, guidance and

promoting the conferences. I also wish to thank all members of programs committees for reviewing the papers. The two program chairs, Dr Li Haizhou (under the assistance of Dr Zhang Min) and Dr Ji Dong Hong worked particularly hard in preparing the programs and the proceedings and Ms Lua Tse Min for assisting in the final typesetting and preparation of the proceedings. I also wish to thank the council members of colips for their supports and funding of the conferences. The travel grants received by the young researchers come from the colips council and Lee Foundation of Singapore jointly. Lee Foundation has been always very supportive to our academic activities. Last but not least, we also wish to thank staff members of the Sijori Hotel who help organize the conference.

Wish all of you the very best!

By

Lua Kim-Teng, Local Chair, O-COCOSDA2003 and PACLIC17
15-08-2003

PACLIC17 Conference Committees

International Co-Chairs		
	Huang Chu-Ren,	Academia Sinica, Taipei, Taiwan
	Akira Ikeya,	Toyo Gakuen University, Tokyo
	Ik-Hwan Lee	Yonsei University, Seoul
	Kawamori Masahito	NTT, Japan
	Byung-Soo Park	Kyung Hee University, Seoul
	Benjamin T'sou	City University of Hong Kong, Hong Kong
Local Chair	Kim Teng LUA	National University of Singapore
Program Committee	Dong Hong Ji (Chair)	Laboratories for Information Technology, Singapore.
	Jing-Shin Chang	Dept. of CSIE, National Chi-Nan Univ., Puli, Nantou, Taiwan
	Hsin-Hsi CHEN	National Taiwan University, Taipei, Taiwan
	Keh-Jiann CHEN	Research Fellow, Institute of Information Science, Academia Sinica, Taipei, Taiwan
	Key-Sun CHOI	KAIST Korea Terminology Research Center for Language and Knowledge Engineering, Daejeon, Korea
	Zhiwei FENG	Institute of Applied Linguistics, Beijing, China
	Yasunari HARADA	Waseda University, Tokyo, Japan
	One-Soon Her	National Chengchi University, Taipei, Taiwan
	Kaoru HORIE	Tohoku University, Japan
	Kiyoshi, ISHIKAWA	Hosei University, Tokyo, Japan
	Beom-mo KANG	Korea University, Korea
	Sue-Jin, KER	Soochow University, Taipei, Taiwan
	Jong-Bok KIM	Kyung Hee University, Seoul, Korea
	Tom Bong Yeung LAI	City University of Hong Kong, Hong Kong
	Chungmin LEE	Seoul National University, Seoul, Korea
	Kiyong LEE	Korea University, Seoul, Korea
	Wei LI	Cymfony Inc., Buffalo, USA
	Robert Wing Pong LUK	Hong Kong Polytechnic University, Hong Kong, China
	Kang Kwong LUKE	University of Hong Kong, Hong Kong
	Yuji MATSUMOTO	Nara Institute of Science and Technology, Nara, Japan
	Tetsuharu MORIYA	Kanazawa University, Kanazawa, Japan
	Jian-Yun NIE	University of Montreal, Montreal, Canada
	Fuji Ren	The University of Tokushima, Tokushima, Japan

Richard SPROAT	AT&T Labs -- Research, Florham Park, NJ, USA
Lily I-wen SU	National TaiwanUniversity, Taipei, Taiwan
Maosong SUN	Tsinghua University, Beijing, China
Shu-Chuan TSENG	Institute of Linguistics, Academia Sinica, Taipei, Taiwan
Jie XU	National University of Singapore, Singapore
Jianqi WANG	Ohio State University, Columbus, USA
Chung-Hsien WU	National Cheng Kung University, Tainan, Taiwan
Shiwen YU	Peking University, Beijing, China

MESSAGE FROM PACLIC 17 & O-COCOSDA 2003 PROGRAM COMMITTEE

It is with great pleasure that we welcome you to Singapore for PACLIC 17 and Oriental COCOSDA 2003. It provides an excellent forum for scientists and engineers from various research areas, including linguistics, natural language processing, and speech processing, etc. to meet and to exchange ideas, to share information and to discuss regional matters on any related research issues.

The call-for-papers this year was met with such enthusiasm that we are able to prepare a technical programme that comprises a large number of high quality papers. For PACLIC 17, we received more than 80 submissions from 13 countries or regions: China, Hong Kong, Taiwan, Japan, Korea, Singapore, Thailand, Vietnam, Mongolia, United states, United Kingdom, France and Australia. For O-COCOSDA, we received over 50 submissions from 9 countries or regions: China, Hong Kong, India, Indonesia, Japan, Korea, Singapore, Thailand and Taiwan. In addition to the 47 papers for PACLIC (36 for oral presentation and 11 for poster) and 39 papers for COCOSDA accepted for presentation, we are honoured to have 4 keynotes speeches given by academic and industrial leaders in the field.

The excellent program offered by PACLIC 17 and Oriental COCOSDA 2003 is a result of the excellent review efforts undertaken by the two program committees and the local organizing committee here in Singapore. We would like to thank the steering committees for their guidance, the program committees for their great reviews, the organizing committees for their hard working and many people who contributed to PACLIC 17 and Oriental COCOSDA 2003 in one way or another. Especially, we would like to thank all the authors who submitted their papers.

On behalf of the program committees, we thank you very much for your tremendous supports in making today's event a success. October is the best to visit Singapore. With the intensive technical program, the open atmosphere and the multi-cultural surroundings, we hope that you can take this excellent opportunity to meet old friends and to make new ones. After SARS, the Lion City is roaring again, we are sure that you will enjoy the city as well as the conference.

Ji Donghong
PACLIC 17 program Committee Chair

LI Haizhou
Oriental COCOSDA Program Committee Chair

Table of Contents

Invited Speeches

Virtual Linked Lexical Knowledge Base for Causality Reasoning 1
Choi Key-Sun

Focus-Marking in Chinese and Malay: A Comparative Perspective 2
Jie Xu

Paper Presentation

Linguistics-1

A Constraint-Based Analysis of Association With Focus in Japanese 16
Yusuke Kubota

On the Event Structure of Indirect Passives in Japanese 28
Naoyuki Ono

An Event-Based Interpretation of Japanese Honorific Constructions Using RRG Operators 38
Akira Ishikawa

Multiple Nominative Constructions in Japanese and Their Theoretical Implications 50
Masahiro Akiyama

NLP-1

Automatic Learning of Stemming Rules for the Indonesian Language 62
Lily Suryana Indradjaja and Stephane Bressan

A Simplified Latent Semantic Indexing Approach for Multi-Linguistic Information Retrieval 69
Liu Yi, Lu Haiming, Lu Zengxiang and Wang Pu

An integrated approach for Chinese word segmentation 80
Guohong Fu and K.K. Luke

Korean Phrase Structure Grammar and Its Implementations into the LKB system 88
Jong-Bok Kim and Jaehyung Yang

Porting Grammars between Typologically Similar Languages: Japanese to Korean 98
Roger Kim, Mary Dalrymple, Ronald M. Kaplan and Tracy Holloway King

Linguistics-2

Mandarin Chinese *Shenme* in Interaction 106
Fuhui Hsieh

The Semantics of Shapes: A Study based on Mandarin *Quan1zi5* (圈子) 115
Cui-xia Weng and Chu-Ren Huang

Stock Markets as Ocean Water: A Corpus-based, Comparative Study of Mandarin Chinese, English and Spanish 124
Siaw-Fong Chung, Kathleen Ahrens and Ya-hui Sung

Subject Positions and Derivational Scope Calculation in Minimalist Syntax: A *Phase*-Based approach 134
Yukiko Ueda

NLP-2

Context-rule Model for POS tagging 146
Yu-Fang Tsai and Keh-Jiann Chen

Chinese Word Segmentation based on Contextual Entropy 152
Jin Hu Huang and David Powers

Topic Segmentation for Short Texts 159
Tao-Hsing Chang and Chia-Honag Lee

Cross-Lingual Text Filtering Based on Text Concepts and KNN 166
Li Shaozi, Su Weifeng, Li Tangqiu and Chen Huowang

Linguistics-3

The Semantics of Onomatopoeic Speech Act Verbs 174
I-Ni Tsai and Chu-Ren Huang

Mandarin Adverbial *Jiu* in Discourse 182
Fuhui Hsieh

A Synchronous Corpus-Based Study of Verb-Noun Fluidity in Chinese 194
Oi Yee Kwong and Benjamin K. Tsou

Non-Monotonic Negativity 204
Sumiyo Nishiguchi

NLP-3

The SVM With Uneven Margins and Chinese Document Categorization 216
Yaoyong Li and John Shawe-Taylor

The development of Tagged Uyghur Corpus 228
Yusup Aibaidula and Kim-Teng Lua

A Vector-Based Algorithm for Chinese Text Classification 235
Luo Changri and He Tingting

A Large-scale Lexical Semantic Knowledge-Base of Chinese 243
Wang Hui and Yu Shiwen

NLP-4

An Effective Combination of Different Order N-grams 251
Sen Zhang and Na Dong

Efficient Methods for Multigram Compound Discovery 257
Wu Horng Jyh Paul, Ng Hong I and Gong Ruibin

Translation Template Learning Based on Hidden Markov Modeling 269
Nguyen Minh Le, Akari Shimazu and Susumu Horiguchi

News-Oriented Keyword Indexing with Maximum Entropy Principle 277
Li Sujian, Wang Houfeng, Yu Shiwen and Xin Chengsheng

NLP-5

Extracting Chinese Multi-Word Units from Large-Scale Balanced Corpus 282
Liu Jianzhou, He Tingting and Liu Xiaohua

A New Sentence Reduction based on Decision Tree Model 290
Nguyen Minh Le and Susumu Horiguchi

Japanese Parser on the basis of the Lexical-FG Formalism and its Evaluation 298
Hiroshi Masuichi, Tomoko Ohkuma, Hiroki Yoshimura and Yasunari Harada

A Statistical Approach to Chinese-to-English Back-Transliteration 310
Chun-Jen Lee, Jason S. Chang and Jyh-Shing Roger Jang

On the Sentence Category Transfer of Action-effect Sentences in Chinese-English Machine Translation 319
Keliang Zhang

Linguistics-4

Modeling verb Order in Complex Multi-Verbal Predicate Constructions 328
Olivia S.-C. Lam and Adams B. Bodomo

The Structure of Spatial Expressions in Saisiyat 339
Pei-Shu Tsai

A Constraint-Based Grammar of Case: To Correctly Predict Case Phrases Occurring without Their Head Verb 351
Hiroki Koga

The Floating of Negative Factors and the Recognition of Semantic Patterns of *Huaiyi* Sentences in Mandrain 362
Xiao Guozheng and Guo Tingting

Poster

Using Mutual Information to Identify New Features for Text documents of Various Domains 372
Guo Zhi Li

Applicability Analysis of Corpus-derived Paraphrases toward Example-based Paraphrasing 380
Kiyonori Ohtake and Kazuhide Yamamoto

A Word Selection Model Based on Lexical Semantic Knowledge in English Generation 392
Chen Yi-Dong, Li Tang-Qiu and Zheng Xu-Ling

Corpus-Based Ontology Learning for Word Sense Disambiguation 399
Sin Jae Kang

On Intra-page and Inter-page Semantic Analysis of Web Pages 408
Jun Wang, Jicheng Wang, Gangshan Wu and Hiroshi Tsuda

Towards a Multi-Objective Corpus for Vietnamese Language 416
Vu Hai Quan, Pham Nam Trung, Nguyen Duc Hoang Ha, Huynh Bao Toan, Le Hoai Bac and Hoang Kiem

Using Zero Anaphora Resolution to Improve Text Categorization 423
Ching-Long Yeh and Yi-Chun Chen

Dependency of Long-Distance Reflexives 431
Hyeran Lee

Voicing Constraint and Segmental-Tonal Neighborhood Effects on Clusters in Thai 441
Rattima Nitisaroj

The treatment of Japanese Focus Particles based on Lexical-Functional Grammar 448
Tomoko Ohkuma, Hiroshi Masuichi, Hiroki Yoshimura and Yasunari Harada

Empty Category and the effect of Teaching in Sentence Processing 456
TakakoKawasaki and Kiyoshi Ishikawa