

# Contents

<b>4</b>	<b>Opening Address</b> Setsuho IKEHATA (President of Tokyo University of Foreign Studies)
<b>6</b>	<b>Center of Usage-Based Linguistic Informatics (UBLI)</b> Yuji KAWAGUCHI (COE Program Leader)
<b>11</b>	<b>Conference overview</b>
<b>12</b>	<b>Featured speakers</b>
<b>13</b>	<b>Location</b>
<b>14</b>	<b>Daily schedule</b>
<b>17</b>	<b>Computer-Assisted Linguistics / Corpus Linguistics</b>
17	Pieter van REENEN, Anke JONGKIND
33	Christian LECLÈRE
47	Francisco MORENO-FERNÁNDEZ
65	Hiroto UEDA
83	Yoshiyuki MUROI
93	Masami MIYAMOTO
109	Takuya NAKAMURA
131	Kanetaka YARIMIZU, Yuji KAWAGUCHI, Masanori ICHIKAWA
<b>151</b>	<b>Applied Linguistics</b>
151	Janet HOLMES
171	David BLOCK
187	Suzuko NISHIHARA
191	Takashi SUZUKI, Koji MATSUMOTO, Mayumi USAMI
205	Rie KIBAYASHI, Kyongbun KIM, Yun XIE, Hironori SEKIZAKI, Sachiko KIYAMA, Hsin-yu SHIH, Eunmi LEE, Koji MATSUMOTO, Mayumi USAMI
219	Yun XIE, Rie KIBAYASHI, Sachiko KIYAMA, Hsin-yu SHIH, Eunmi LEE, Kyonbun KIM, Koji MATSUMOTO, Mayumi USAMI
235	Mark PETERSON
243	Malcolm H. FIELD
261	Ai FUJIWARA, Naoyuki NAGANUMA, Tomoko WADA, Koji SHIBANO
273	Kazuya ABE
281	Tsutomu KIGOSHI
291	Kentaro YUKI

# Opening Address

*Setsuho IKEHATA (President, Tokyo University of Foreign Studies)*

The 21st Century COE Program launched by the Ministry of Education, Sports, Culture, Science and Technology in 2002 is a center of research and education in all academic fields, a Center Of Excellence, that sets the highest standards for the world. Its goals are to raise the level of research in our country's universities and develop creative academic minds to lead the world. Tokyo University of Foreign Studies (TUFS) submitted applications for research projects in two of the selected programs—the Humanities and Interdisciplinary/Compound/New Sphere fields. We got wonderful results; both projects were selected. We are tremendously pleased and encouraged by this high evaluation of the unique research and educational potential of our Graduate School of Area and Culture Studies.

At the University, outstanding experts from every area of the world—Europe, North and South America, Oceania, Asia, and Africa—collaborate on education and research in a wide range of academic fields including linguistics, literature, history, philosophy, cultural anthropology, sociology, political science, and economics. In this sense, although we are a single-faculty university, we have attained an extremely consistent interdisciplinary and comprehensive approach. In an age that emphasizes the global community, it is certainly desirable for us to maximize and further develop this unique strength in both education and research.

A strong foundation in foreign languages is vital to area and culture studies. The University engages in education and research in upwards of 50 languages and cultures and societies in every part of the world that contributes to cross-cultural understanding and the development of people capable of contributing to the actualization of a harmonious global community. In addition, a double-major system that requires students to specialize in both a language and a discipline-related course of study enables the University to send graduates into society equipped with a high degree of language competence and a deep knowledge of world cultures and societies.

The new Fuchu campus has greatly improved facilities with an advanced computing network. The University prides itself on its information literacy and the number of computers on campus, which ranks at the top level among liberal arts universities in the country. We plan to employ this wonderful information infrastructure as a foundation from which to draw on the strengths of multimedia and the Internet and utilize a variety of methods to promote the most advanced language education possible. And we are leveraging the University's unique characteristics to develop comparative studies in Japanese and other languages.

The University's Usage-Based Linguistic Informatics project, selected by the 21st Century COE Program, is the concrete manifestation of our plans for the future, which I have just mentioned. The implementation team members are committed to this future vision and vigorously engaged in the project. It is my fervent desire that they will produce rewarding results. It is the intention of everyone at Tokyo University of Foreign Studies to combine our wisdom in a concerted effort to do our utmost to make a success of the 21st Century COE Program.

In closing, let me welcome the distinguished authorities in a variety of fields from the Netherlands, France,

Germany, Spain, New Zealand, and England who we have invited to be our guest speakers. Thank you so much for adjusting your busy work schedules and for coming such a long way to participate in this international conference. As President of Tokyo University of Foreign Studies, I also wish to express my deep gratitude to each of our guests from research institutions throughout Japan who have taken the time to attend the conference. I hope it will be a great success and a productive and rewarding experience for all of you.

Tokyo, October 15, 2003



# Center of Usage-Based Linguistic Informatics (UBLI)

Yuji KAWAGUCHI (COE Program Leader)

Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies (TUFS)

## 1. Linguistic Informatics

It is widely believed that linguistic theories and computer sciences have much influenced foreign language education, while the collaboration of these three domains has not brought about new scientific results. The present program will meet such a scientific need. An overall integration of Theoretical and Applied Linguistics will be realized on the basis of Computer Sciences. We have named this new synthetic field *Linguistic Informatics*. When we hear this name for the first time, we may take it for a branch of natural sciences. However, since our language represents a system of information, Linguistics itself constitutes, in a broad sense, a part of Informatics. In the following lines, the limitation of space will oblige me to explain only the essence of this *21st Century COE (Center of Excellence) Program*.

## 2. Organization and Research Projects

The present COE program is directed by the following supervisors: Susumu ZAIMA (German Linguistics), Toshihiro TAKAGAKI (Spanish Linguistics), Yoichiro TSURUGA (French Linguistics), Kohji SHIBANO (Computer Sciences), Makoto MINEGISHI (Theoretical Linguistics), Mayumi USAMI (Applied Linguistics) and Yuji KAWAGUCHI (French and Turkish Linguistics). In the academic year 2003, the following research projects are undertaken respectively in three scientific fields.

### Research Projects in Academic Year 2003

LINGUISTICS : Corpus Analysis, Syntax and Prosody
Responsibles: Y. KAWAGUCHI, F. KAWAMURA, T. MIYAKE, H. NAKAZAWA, I. SHOHO, K. SOHMIYA, Y. TSURUGA, T. TAKAGAKI , K. URATA
APPLIED LINGUISTICS : Discourse Analysis, Second Language Acquisition, Evaluation of TUFS Language Modules
Responsibles: M. NEGISHI, T. UMINO, M. USAMI, A. YOSHITOMI
COMPUTER SCIENCES: E-learning, Natural Language Processing
Responsibles: CH. LIN, H. SANO

In principle, these projects are considered as fundamental researches for the development of *TUFS Language Modules*, which are the very fruits of Linguistic Informatics and the significant scientific contribution of this COE.

### 3. TUFS Language Modules

#### 3.1. Cohabitation of Natural Language and Machine Language

Our main objective is to innovate foreign language education by developing superior educational material and transmitting it through the Internet. At present, the following 17 languages are covered in TUFS Language Modules.

TUFS Language Modules Editors of Pronunciation and Dialogue Modules	
English	H. SAITO, A. YOSHITOMI
German	T. NARITA, A. MASAKI
French	Y. KAWAGUCHI, A. MIZUBAYASHI
Spanish	S. KAWAKAMI
Portuguese	N. KUROSAWA, CH. TAKEDA
Russian	H. NAKAZAWA
Chinese	K. HIRAI, N. MIYAKE
Korean	I. CHO, K. IKARASHI
Mongolian	Y. SAITO, R. NUKUSHINA
Indonesian	M. FURIHATA
Filipino	T. MORIGUCHI, M. YAMASHITA
Lao	R. SUZUKI
Cambodian	H. UEDA, T. OKADA
Vietnamese	Y. UNE, H. TAHARA
Arabic	R. RATCLIFFE
Turkish	M. SUGAHARA
Japanese	Y. SATO, T. UMINO

This is a large-scale project that includes more than 100 researchers and graduate students. In TUFS Language Modules, the multilingual language learning system would be one of the main characteristics. In fact, we teach more than 40 different languages at TUFS. But the novelty of TUFS Language Modules lies in another fact. For example, 17 languages are described in unicode (UTF-8), and in our system, HTML, a basic language of World Wide Web (WWW), is correlated with XML, which was first invented in 1998 and has recently begun to be applied in WWW.

Improving education for graduate students is also one of the important aims of this COE. Graduate students create basically the preliminary materials for the modules. Through this research activity, they



will gain the knowledge not only of Linguistics and Applied Linguistics, but also of Computer Sciences. In this way, the program will foster new types of linguistic researchers who have full knowledge of Theoretical and Applied Linguistics and can manipulate a computer-assisted language learning system.

### 3.2. *Modularized View of Language*

With the appearance of the Internet, we have become conscious of the omnipresence of information, that is what we call ubiquity of information. On the other hand, WWW gives us an opportunity to think over again how and what the information should be. On WWW, theoretically speaking, infinite ordering and combination of information are possible through their mutual linkages. In TUFs Language Modules, we set our way of thinking free from a traditional view of language and adopt a *modularized view of language*. Each language unit is composed of four relatively independent modules, i.e. pronunciation, dialogue, grammar and vocabulary modules. The idea of module components allows learners and teachers to learn and teach the language in question from whichever part of the modules and in whatever order.

### 3.3. *Cross-Linguistic Syllabus*

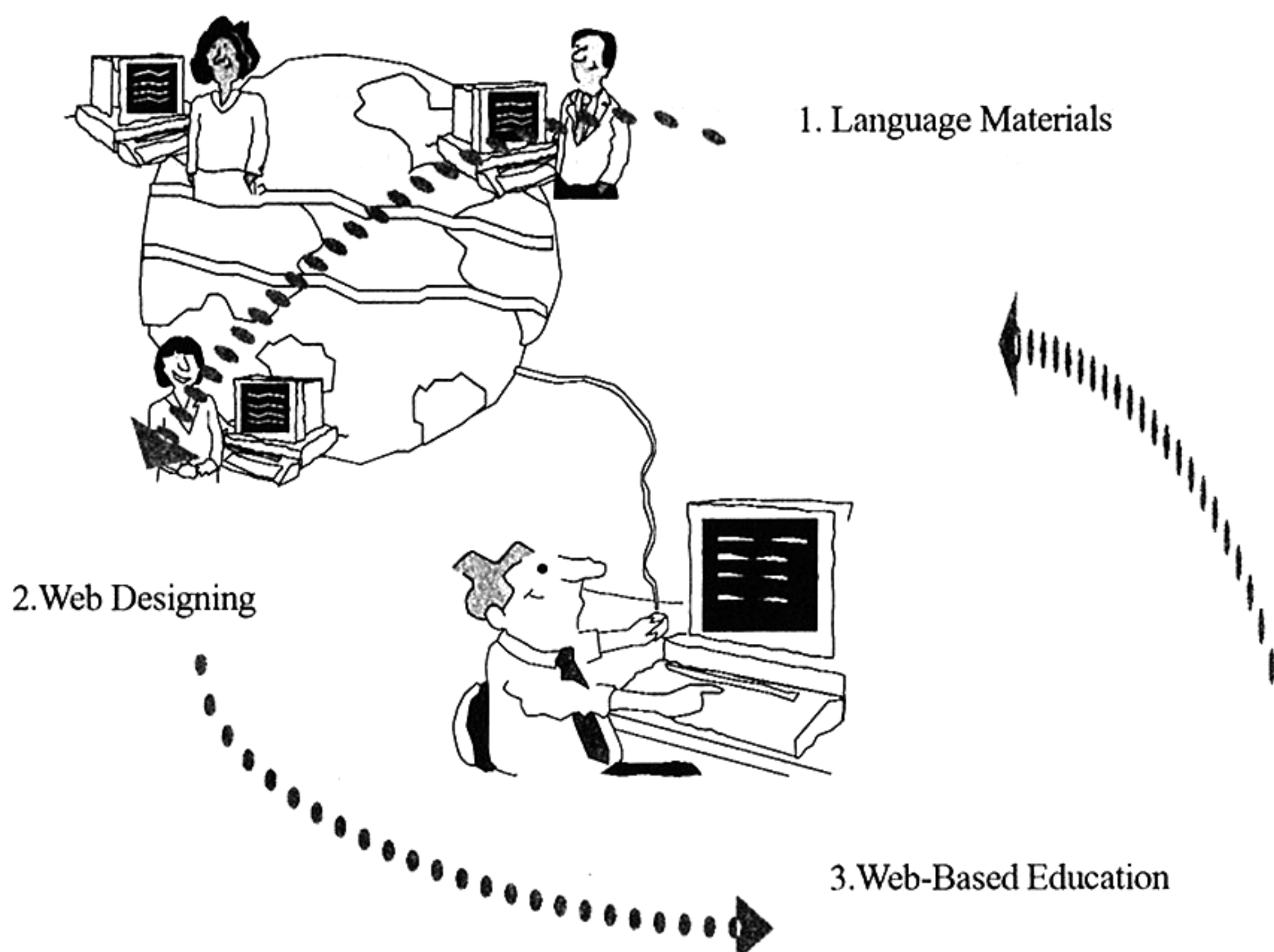
More freedom than ever will be promised to learners and teachers by these modules. However, a common measure is indispensable for the evaluation of language learning and education. In this sense, the evaluation of modules is very important for this COE program. As each module is designed independently to some extent, one may evaluate it individually. But as far as educational contents and goals are concerned, a more or less loose unity has been realized by adopting a common syllabus design for 17 different languages, so that in addition to a traditional analysis of learners' idiosyncratic characters, one can make an interesting contrastive analysis of individual or universal characteristics of second language acquisition (SLA) through 17 different languages. *Cross-linguistic syllabus* is therefore regarded as an innovation in this web-based language education system.

### 3.4. *Linguistic Usage*

The process of developing TUFs Language Modules is as follows: 1. Making language materials; 2. Implementation on WWW; and 3. Web-Based Language Education. Thus, the first step consists in making language materials appropriate for language modules. What kind of language materials must we furnish? We suppose that these language materials should be "usage-based". The key concept here is *linguistic usage*. Then, what in the world does this usage mean? The term is highly polysemous. Some researchers claim that linguistic usage will become explicit only through quantitative analysis of an enormous corpus. Others declare that usage should be fixed in mutual speech acts between a speaker and a hearer. Others suppose that linguistic usage is related to our cognition, for our linguistic knowledge will be accumulated through the encounter with new linguistic usages. We also find researchers who will insist on the interaction of both linguistic and extra-linguistic aspects of linguistic usage. In fact, the definition of usage is not at all unanimous among linguists. TUFs Language Modules give us an opportunity to reconsider the significance of usage for linguistic research and language education. Therefore, isn't it necessary that every researcher and graduate student involved in this program should keep their own opinion on the concept of linguistic usage?



At the end of the year 2003, the pronunciation and dialogue modules will be available in Japanese on the Internet. The development of the grammar and vocabulary modules is underway.



#### 4. First International Conference on Linguistic Informatics

Immediately after the selection of this COE program by the Ministry of Education, Culture, Sports, Science and Technology, we began to prepare for the present conference. At the end of 2002, the outline was fixed. On December the 13<sup>th</sup> and 14<sup>th</sup>, the first International Conference on Linguistic Informatics is planned to be held at Tokyo University of Foreign Studies. The conference has three different sessions: 1. *Computer-Assisted Linguistics*; 2. *Corpus Linguistics*; and 3. *Applied Linguistics*.

It is a great honour for me to organize this international conference, because we have many guest speakers not only from other universities in Japan, but also from all over the world. We have also many graduate students, mostly PhD candidates, who read papers in this conference. As opposed to normal conferences, we prepare prepublished Proceedings before the conference. This conference covers such large scientific fields, i.e. Computer Linguistics, Philology, Dialectology, Corpus Linguistics, Discourse Pragmatics, Applied Linguistics and E-Learning, so that without assistance of prepublished papers, our audience will not be able to understand the essence of contributions and to follow what they are discussing.

However, it would be better to notice that, mainly because of the limitation of time, this volume remains a provisional version of the Proceedings to be published. You may find some spelling errors or some expressions to be more sophisticated from the scientific viewpoint. In spite of these reserves, we hope that this prepublication will help the audience understand the contributions. Through this conference, we expect to know the state of art of Linguistic Informatics and the problems which this new field will have to solve. We hope that this synthesis of different scientific fields is fruitful and gives us some insights into

a future vision of this new science.

Finally, I'd like to express my gratitude towards my colleagues and graduate students of TUFS, and many collaborators of this COE program.

Tokyo, September 30, 2003

cf. Tufs Language Modules: <http://www.coelang.tufs.ac.jp/modules/> (in Japanese)

Usage-Based Linguistic Informatics: [http:// www.coelang.tufs.ac.jp/](http://www.coelang.tufs.ac.jp/) (in Japanese and English)



**第 1 回言語情報学国際会議予稿集**

**Proceedings of the First International Conference on Linguistic Informatics**

**言語情報学 - 現状と未来 -**

Linguistic Informatics - State of the Art and the Future -

©2003 Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies (TUFS)  
3-11-1 Asahi-cho, Fuchu City, Tokyo 183-8534 Japan

All rights reserved. No part of this work may be reproduced or  
Utilized in any form or any means, electric or Mechanical,  
including illustrations without the prior written permission of Graduate School of  
Area and Culture Studies, Tokyo University of Foreign Studies (TUFS) .

Editor : Yuji KAWAGUCHI Susumu ZAIMA  
Toshihiro TAKAGAKI Kohji SHIBANO Mayumi USAMI

Project Coordination: Naomi Tanoura  
Design Coordination : Noriko Sakai  
Logo-Design and Cover Design : Miwa Hirose  
DTP: DesignEXchange  
English proofreading: JEX  
Printing & binding: Hikari Shashin Printing Co., Ltd.