



Multilingual Information Processing for Digital Libraries

Akira Maeda

Department of Computer Science
Ritsumeikan University



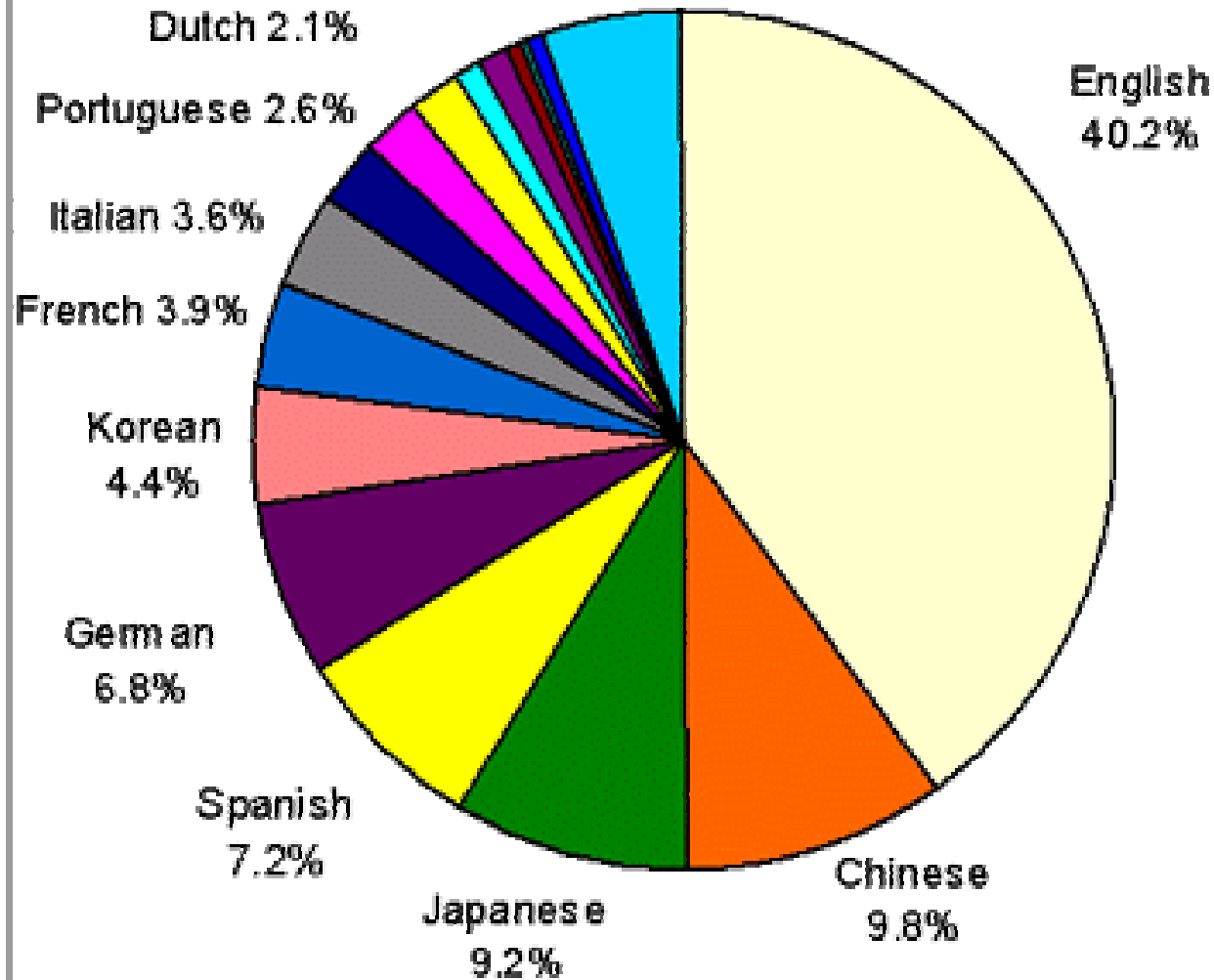
Background

- Increasing popularity of the Internet in various areas in the world
 - Languages used for Web documents are expanded from English to others
- In a digital library (DL), multilingual information processing is essential
 - Even a small library has multilingual materials
 - Accessed from all over the world

Online Language Populations

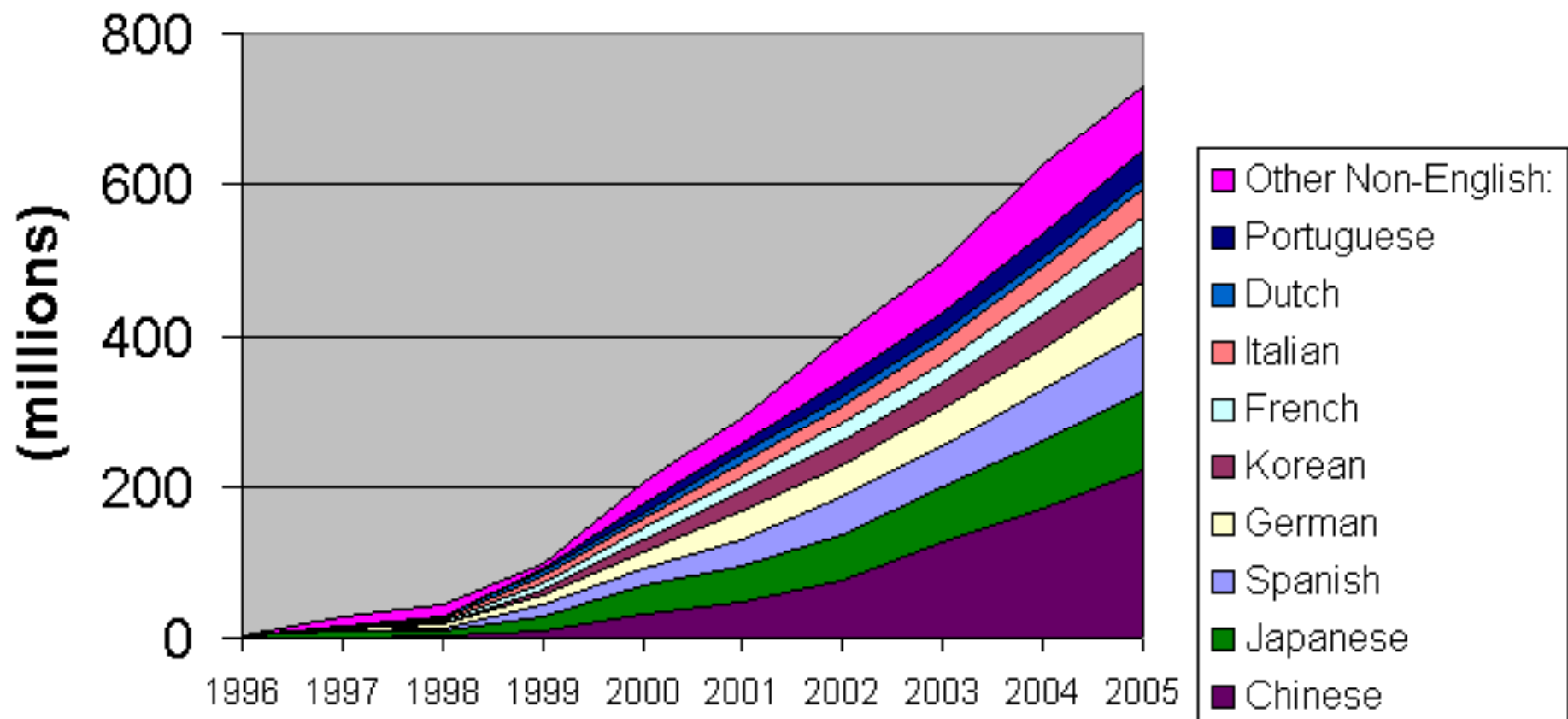
Total: 561 Million

(March, 2002)



Source:
Global Reach

Evolution of non-English-speaking online population



Source: Global Reach



Problems

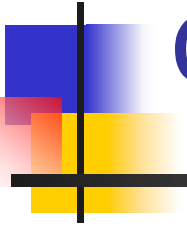
- Many unsolved problems in order to handle multilingual documents in a DL
- From the user's point of view, fundamental functions for the general use of a DL are: **display**, **input**, and **retrieval**
- From the system's point of view, digital documents often lack information of the **encoding and the language**
 - In web documents, `charset` parameter is not always attached
 - incorrect display on browsers
 - inaccurate indexing on search engines



Proposed solutions

- 1. Display and input functions for multilingual text** which does not depend on installed fonts and input methods
- 2. Automatic identification of languages and coding systems of Web documents** based on statistics and heuristics
- 3. Cross-languages IR technique** which is suitable for documents in diverse domains

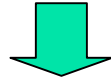
MHTML: Display and input functions for multilingual documents





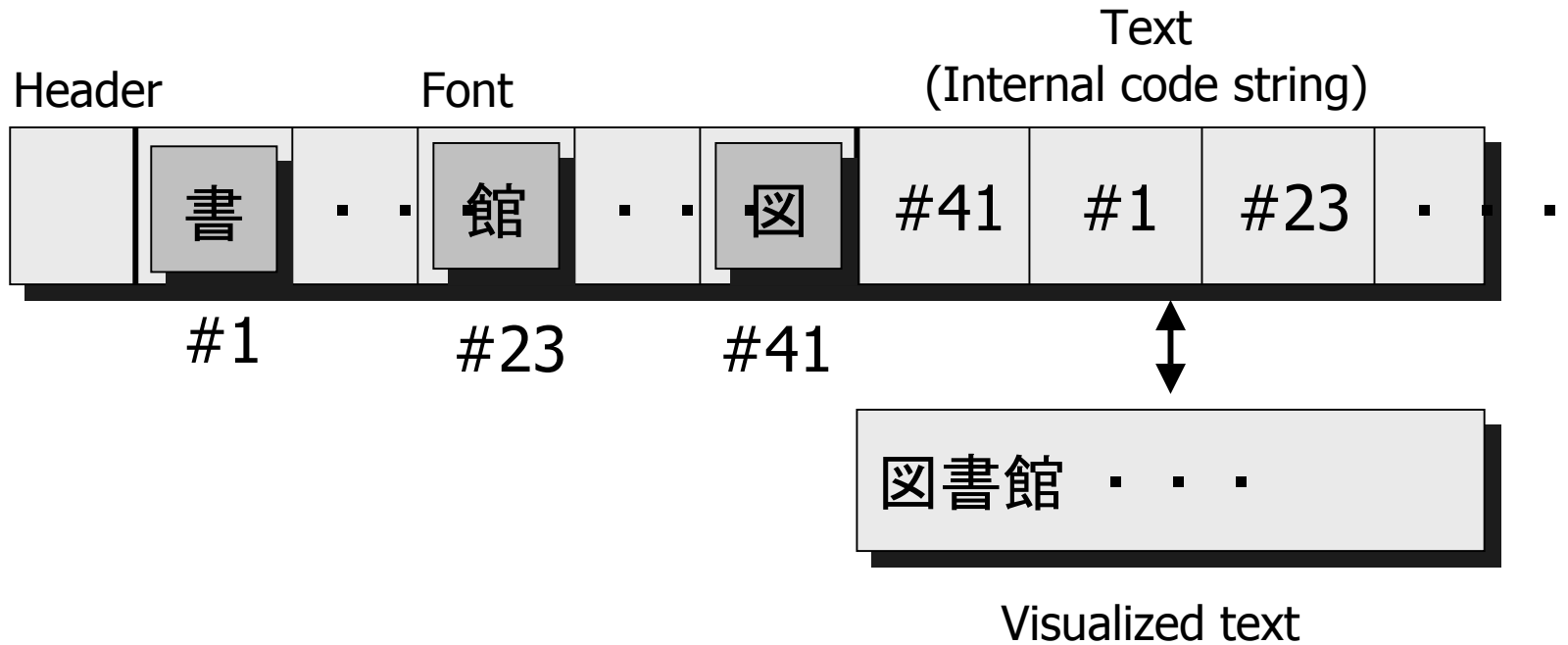
Motivation

- To enable display/input of foreign language text on an off-the-shelf Web browser
 - Usually, the user have to install fonts or input methods for foreign languages
 - Difficult for novice users
 - Impractical for client PCs in public use

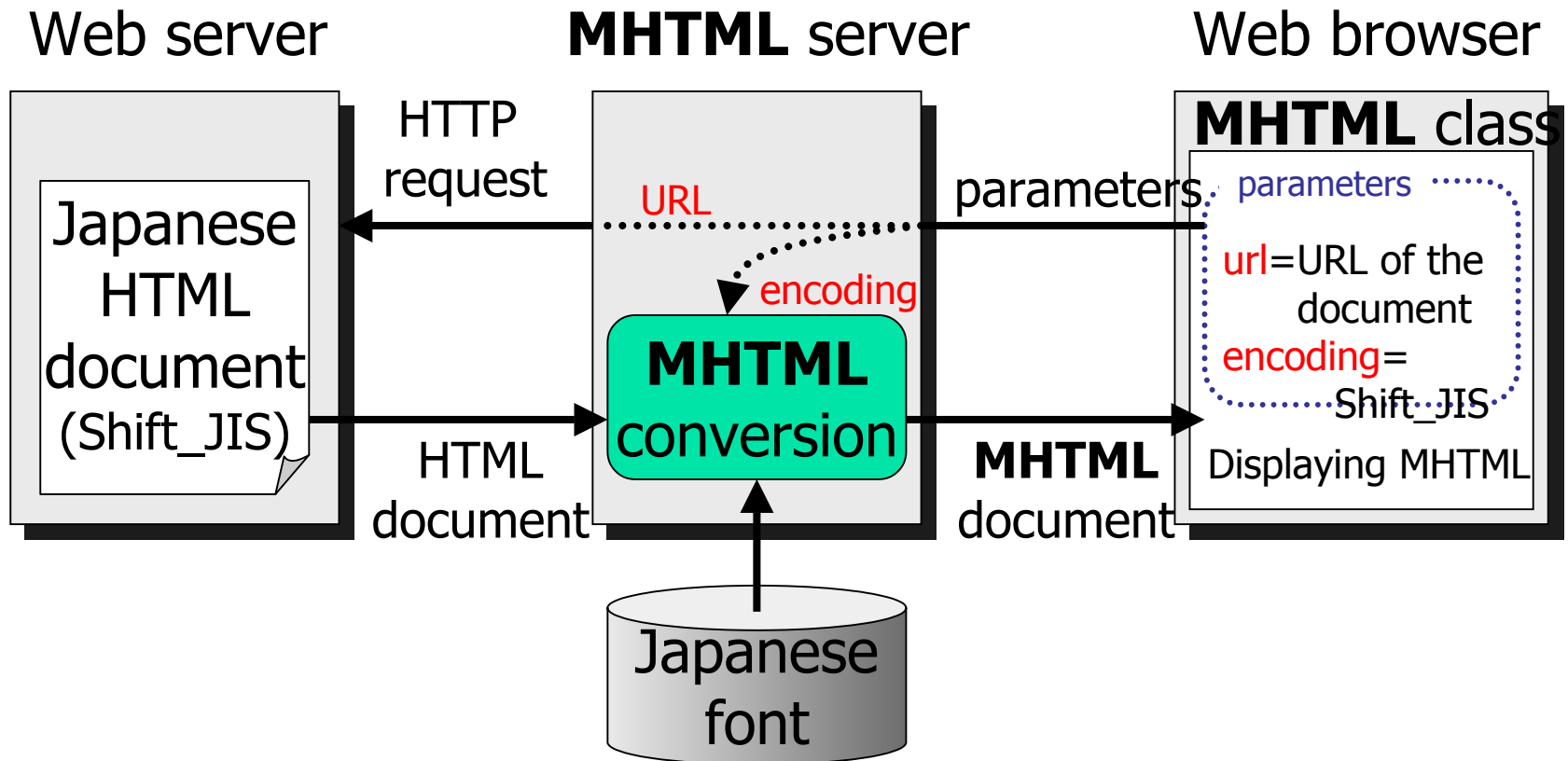


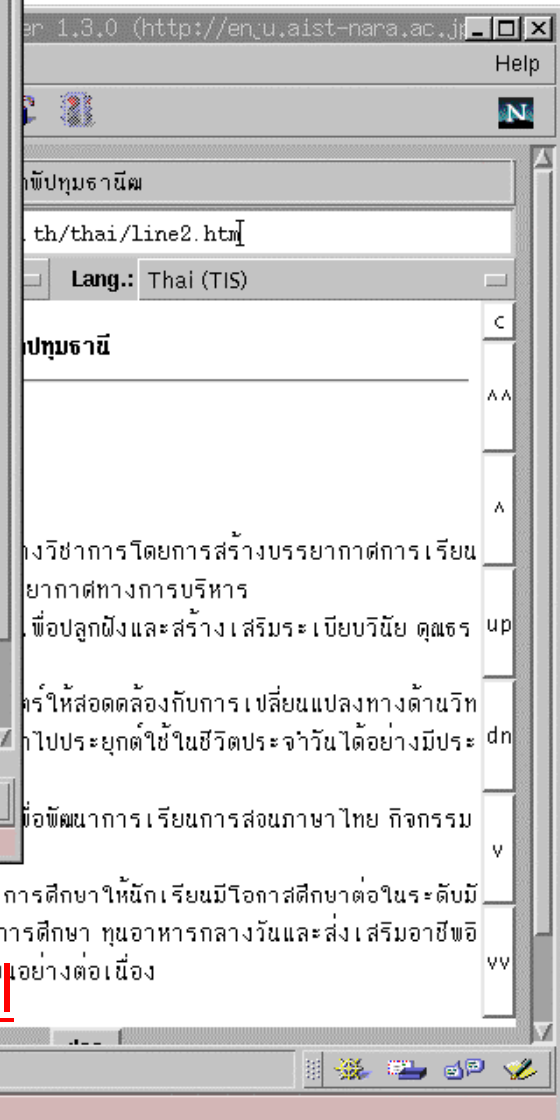
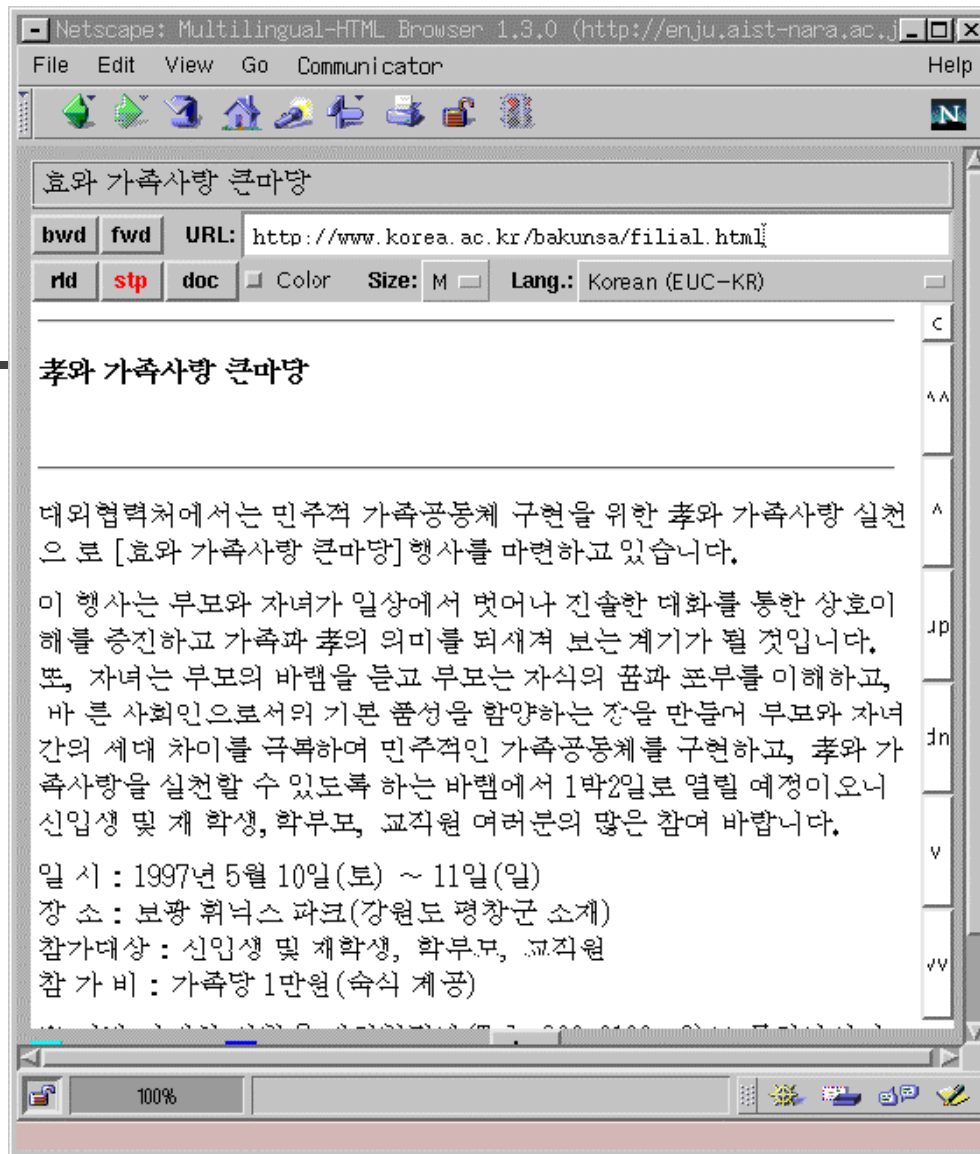
- A method which does not require installing fonts or input methods on the client
- **MHTML**: Multilingual-HTML

Structure of MHTML document



Displaying a document using MHTML






<http://mhtml.ulis.ac.jp/>

<http://www.vt.edu/misc/publish/mhtml.html>

Digital library of multilingual old tales



The screenshot shows a Microsoft Internet Explorer browser window displaying a website titled "The Japanese Old Tales Site". The address bar shows the URL: oshita.4+02-jp_JP/momo.html03-momo.Nobuteru_Shiroshita.4+10-ko_KO/momo.html06-momo.Nobuteru_Shiroshita.4. The page content is organized into three sections, each with a title, a text block, and a set of navigation buttons (c, -, ^, up, dn, v, vv) on the right side.

Monotaro
Once upon a time an old man and woman lived in the mount ains. Everyday the old man went to the mountain and collecte d firewood, while the old woman went to the river and did th e laundry. One day, she was doing the washing when a big pea ch came floating down the river towards her. As it was a big

momotarou
むかしむかし、あるところにおじいさんとおばあさんがいま した。いつも、おじいさんは山へしばかりに、おばあさんは 川へせんたくに行っていました。ある日おばあさんがせんたく をしていると大きなももがどんぶらこどんぶらことながれて きました。おばあさんはそのももを見て、おじいさんにおい

모모타로
옛날옛날 어떤 곳에 할아버지와 할머니가 살고 있었습니다. 언 계나 할아버지는 뽕감을 구하러 산에, 할머니는 팥리를 하러 강 으로 갔습니다. 어느날 할머니가 팥리를 하고 있는데, 커다란 복숭아가 뚝뚝 떠내려 왔습니다. 할머니는 그 복숭아를 보고, 할아버지께 드리려고 집으로 가지고 왔습니다. 점심때가 되어

At the bottom of the page, there is a "Home" link, a row of flags representing various languages, and a message in French, Japanese, and English: "Merci de faire connaitre vos commentaires et questions ! 感想やコメントのある方は、メールを送って下さい。 Thank you for sending comments and questions !". Below this is the email address myria@ulis.ac.jp and a small illustration of a landscape with a house and trees.

<http://www.dl.ulis.ac.jp/oldtales/>



Summary

- Multilingual text display/input functions without installing anything
- Simple and efficient
- No need to modify existing HTML documents
- Useful for minor languages
- Useful for displaying characters that are not defined in standard character set

Automatic identification of coding systems and languages of documents



Target languages and coding systems

Coding system	Language	Unit	Character range
ISO-2022-JP	Japanese	7	33-126
ISO-2022-CN	Chinese	7	33-126
ISO-2022-KR	Korean	7	33-126
Shift_JIS	Japanese	8	33-252
EUC-JP	Japanese	8	33-126, 142-254
GB2312	Chinese (simplified)	8	33-126, 161-254
Big5	Chinese (traditional)	8	33-126, 161-254
EUC-KR	Korean	8	33-126, 142-254
ISO-8859-1	European languages	8	33-126, 161-254



Proposed identification method

- **7-bit coding systems**

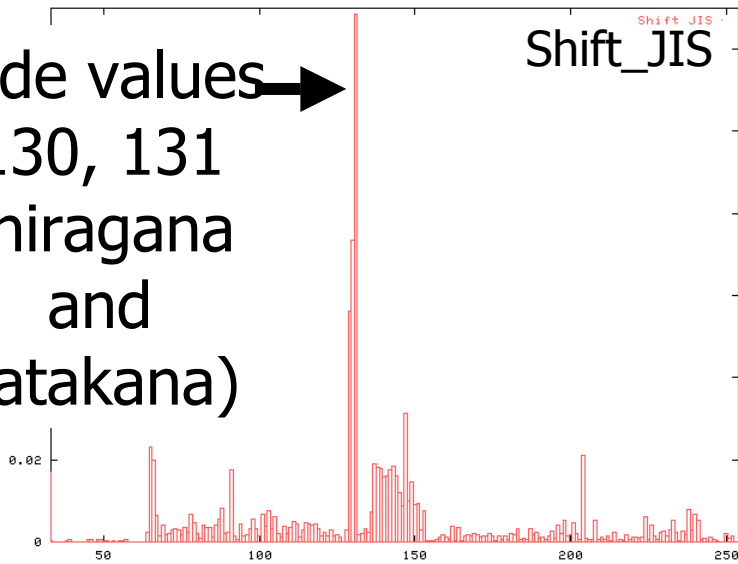
- can be distinguished from 8-bit codes by checking MSB (Most Significant Bit)
- Subsets of ISO-2022 can be distinguished by escape sequences

- **8-bit coding systems**

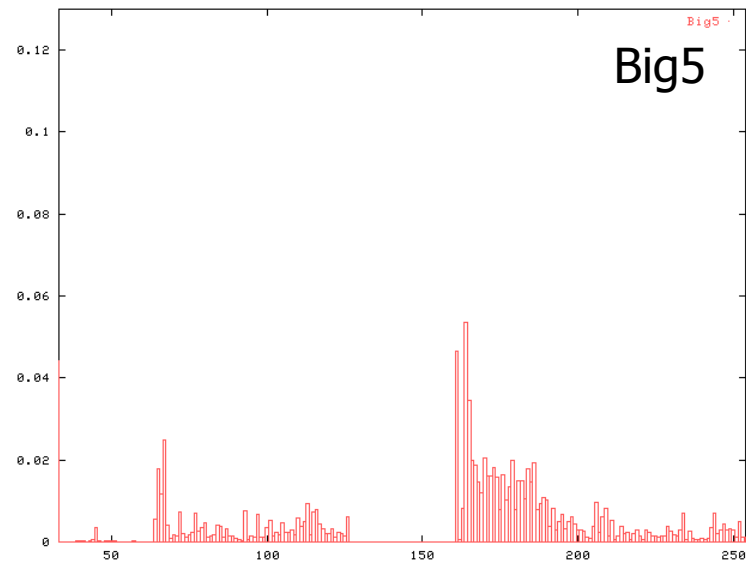
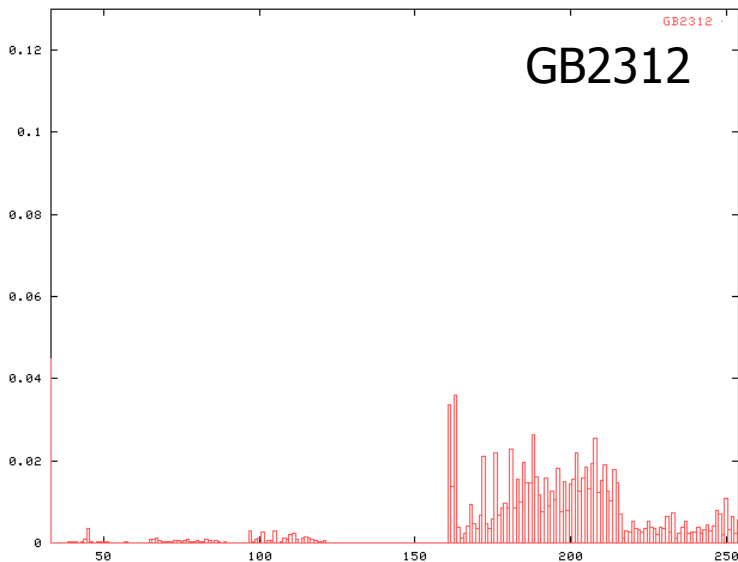
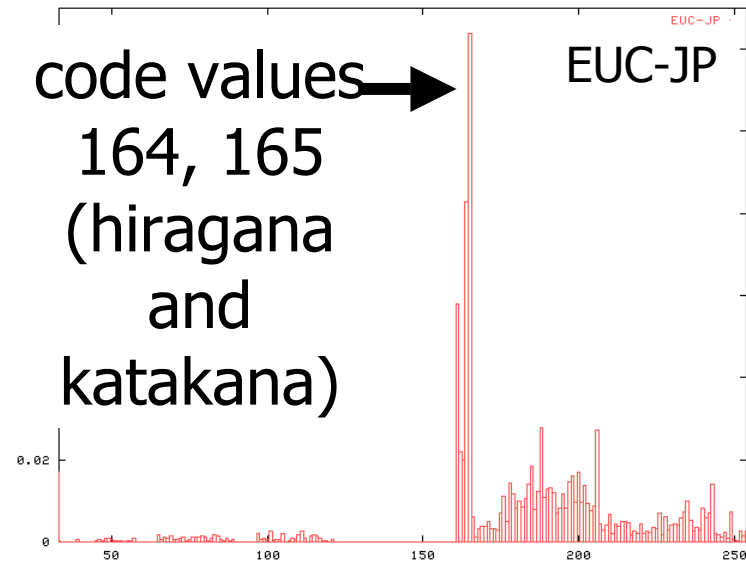
- identified by analyzing the distributions of character codes

One byte code distributions

code values
130, 131
(hiragana
and
katakana)

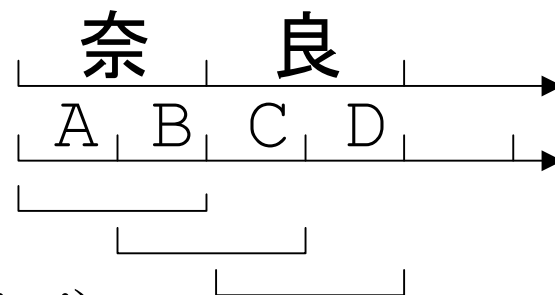


code values
164, 165
(hiragana
and
katakana)



Vector-distance method (consecutive two bytes unit)

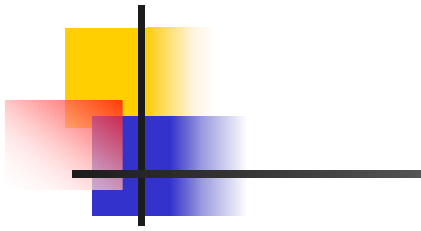
- Takes the connection of consecutive characters into account



$$\cos D'_c = \frac{\sum_i \sum_j freq_c(i, j) freq_d(i, j)}{\sqrt{\sum_i \sum_j freq_c(i, j)^2} \sqrt{\sum_i \sum_j freq_d(i, j)^2}}$$

($i = 32, 65 \dots 90, 97 \dots 122, 128 \dots 255$)

Results of the Vector-distance method



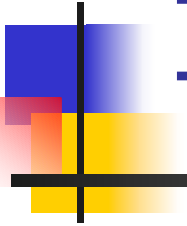
Coding system (language)	Correct rate (%)	
	1 byte	2 bytes
Shift_JIS (Japanese)	99.8	100.0
EUC-JP (Japanese)	99.2	100.0
GB2312 (Chinese)	100.0	100.0
Big5 (Chinese)	100.0	100.0
EUC-KR (Korean)	100.0	100.0
ISO-8859-1 (English)	90.9	98.9
ISO-8859-1 (German)	99.4	100.0
ISO-8859-1 (French)	95.1	99.9
ISO-8859-1 (Italian)	98.2	100.0
ISO-8859-1 (Spanish)	90.7	99.9
ISO-8859-1 (Portuguese)	92.8	100.0
ISO-8859-1 (Danish)	69.8	92.6
ISO-8859-1 (Norwegian)	70.1	91.5
ISO-8859-1 (Swedish)	94.9	99.7
Avg.	92.9	98.7



Summary

- Proposed an identification method of coding systems and languages of Web documents
- 98% average correct rate for 12 languages and 10 coding systems
- The method does not require discriminating the boundaries of characters for Asian languages

Query Term Disambiguation for Cross-Language Information Retrieval





Cross-Language Information Retrieval (CLIR)

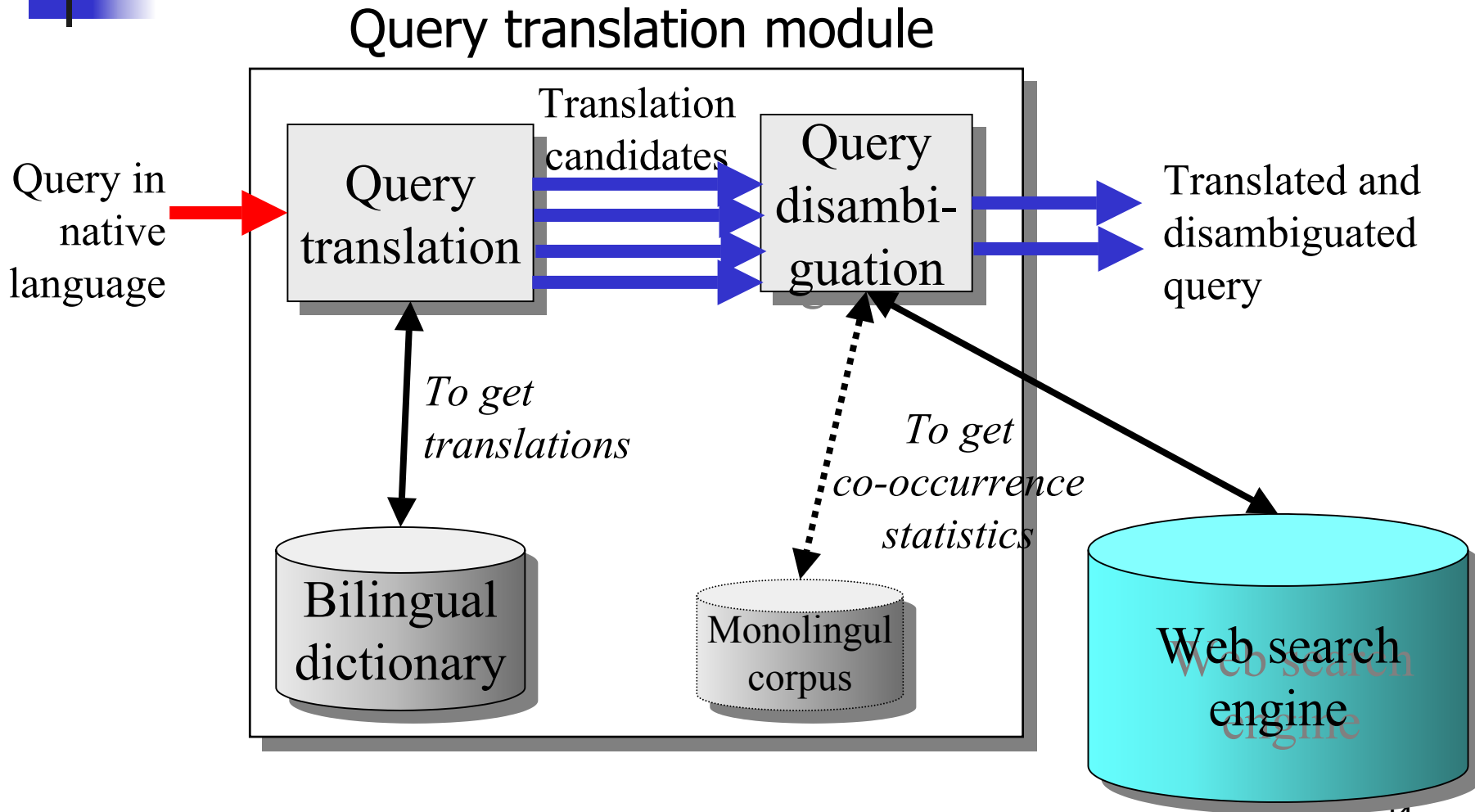
- A technique to retrieve documents written in a certain language using a query written in another language
- Who needs CLIR?
 - When users can read several languages
 - Eliminate multiple queries
 - Query in most fluent language
 - Monolingual users can also benefit
 - If translations can be provided
 - Inexpensive Machine Translation software



Approaches to CLIR

- Translation of target collection
 - Has the advantage of utilizing existing MT software
 - Not suitable for multilingual, large-scale, and frequently-updated Web collection
- Translation of user's query
 - Translated queries can simply be fed into existing monolingual search engines
 - Simple dictionary translation introduces **ambiguity**
 - bank : 銀行 (bank to deposit), 堤防 (dike), 土手 (embankment), 川岸 (riverside) ...

Flow of query translation

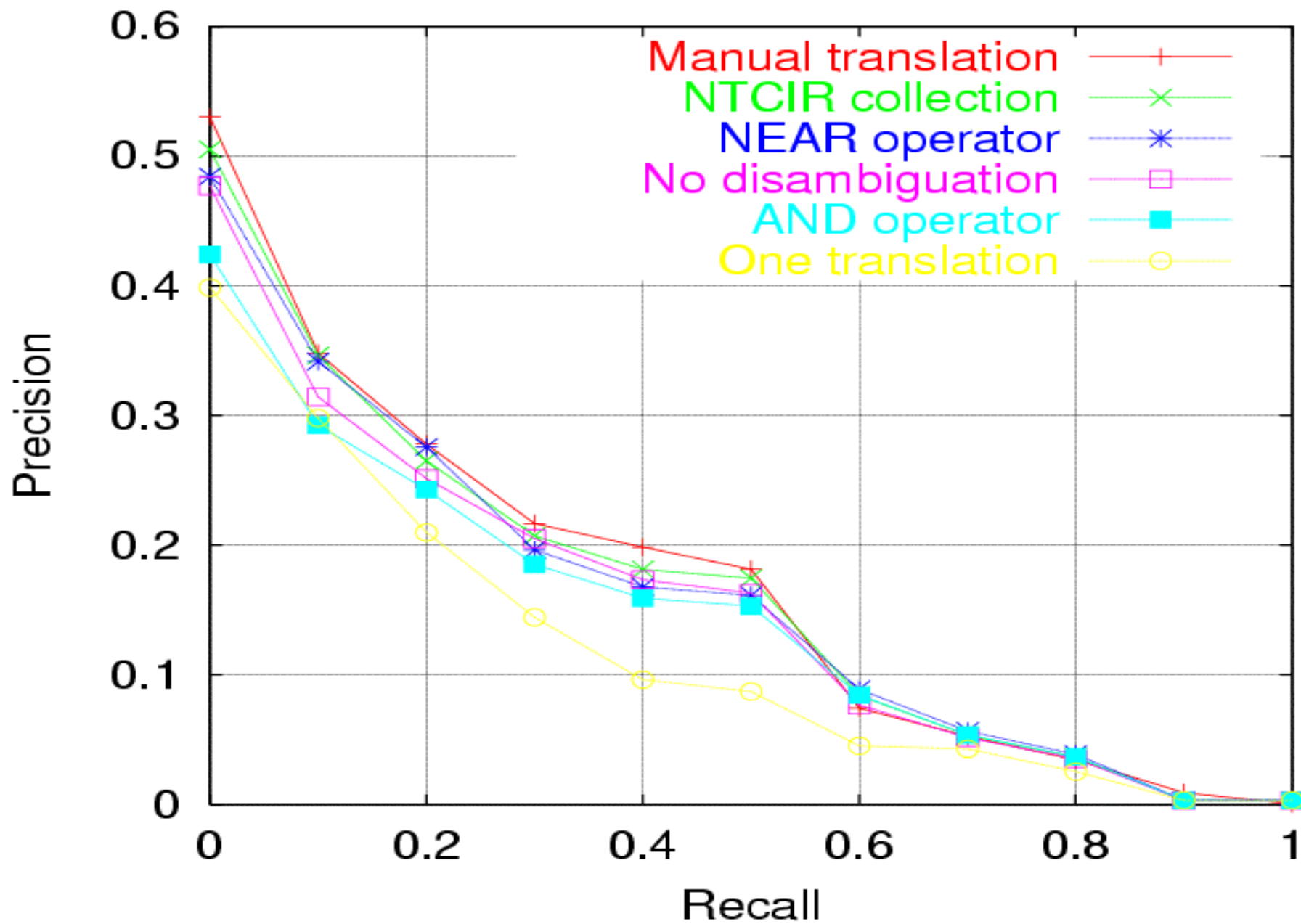




Mutual Information

w1	w2	$COT_{MI}(w1, w2)$
database	multimedia	3.37
database	transaction	2.41
database	relational	2.01
database	chair	-2.96
database	soul	-3.43
database	iron	-4.62

Experiment (MI: n words COT)





Summary

- CLIR method for DL documents in diverse domains
 - utilizing a Web search engine as a corpus
- Effective for very short queries
- Can easily be extended to other languages



Conclusions

- Proposed some solutions to the problem of multilingual information processing for DL
 - **Display, input, and retrieval**