

# 次世代インターフェースとしての 多言語コンシェルジュの研究開発

町田和彦

東京外国語大学

アジア・アフリカ言語文化研究所

# 多言語コンシェルジュ

- 多言語・多文字で表現された情報資源に容易にアクセスするためのサービスとそのヒューマンインターフェースの総称。
- 言語や文字の壁を意識しないで利用できる環境を提供し、言語バリアを取り除く近未来の中核的システム

## 1) 言語コンポーネント

- 文字列(つづり)の標準化
- 文字コード列の正規化

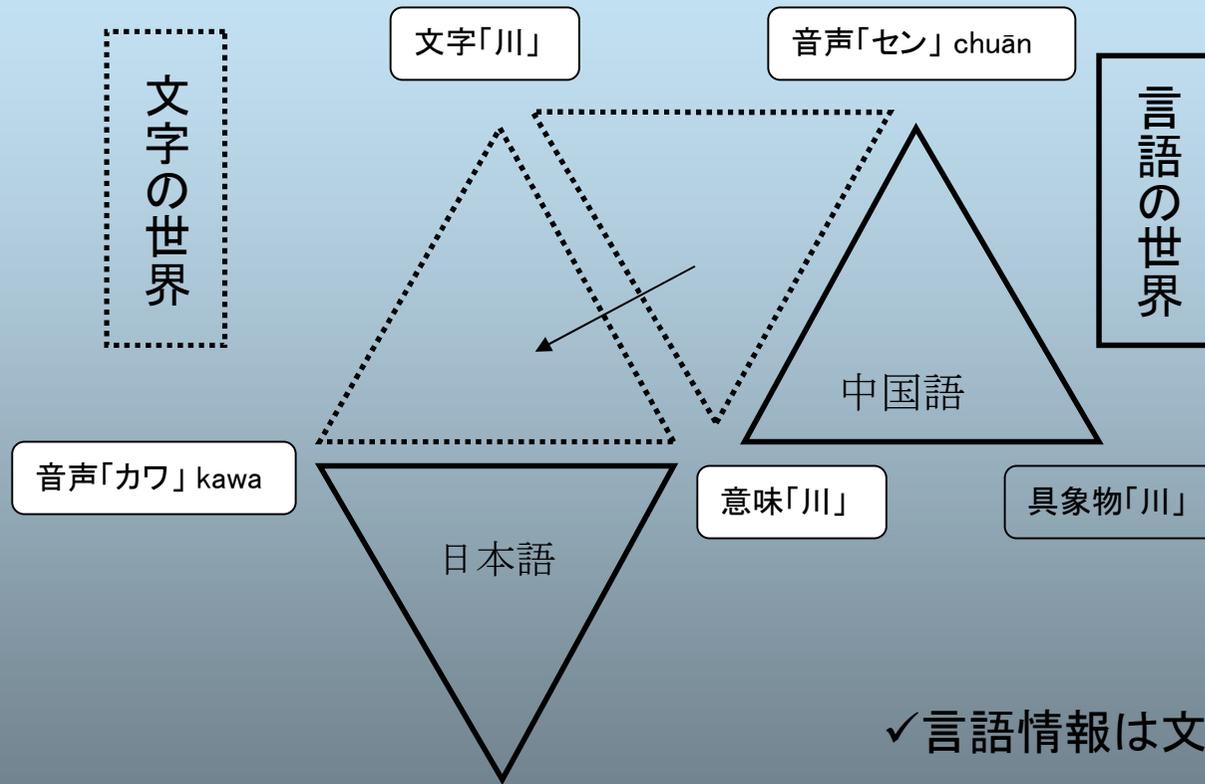
## 2) 多文字環境に適したInput Method

- メタスクリプト(metascript)
- サーバ/クライアント型のモデル実装

# 研究対象：4文字系と24の言語

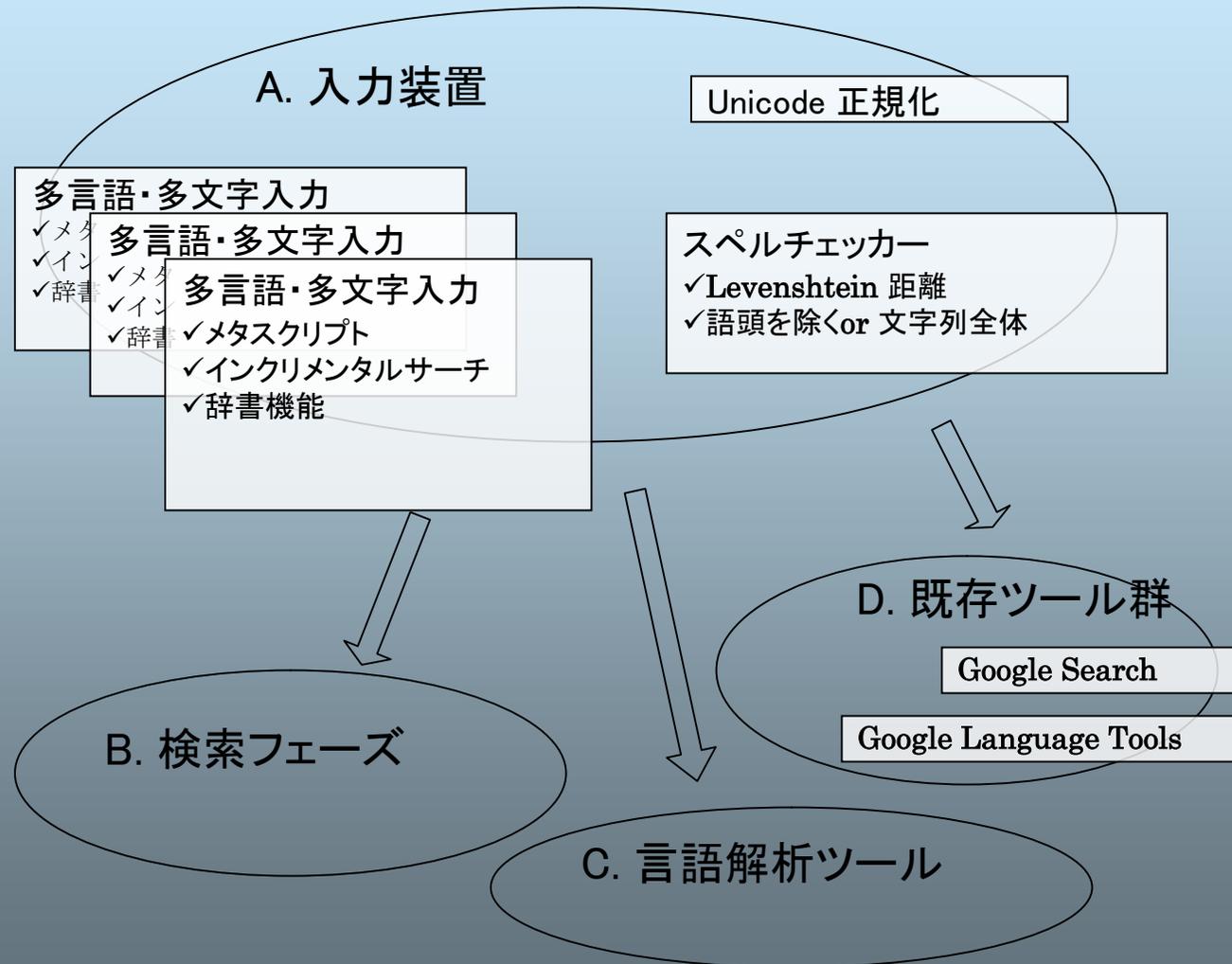
- ラテン文字（英語、ドイツ語、フランス語、スペイン語、イタリア語、ポルトガル語、インドネシア語、ベトナム語、トルコ語、ロシア語）
- アラビア系文字（アラビア語、ウルドゥー語、ペルシヤ語、ウイグル語）
- インド系文字（ヒンディー語、パンジャービー語、シンハラ語、ビルマ語、タイ語、カンボジア語、チベット語、朝鮮語）
- 漢字（中国語、日本語）

# なぜ文字？ 文字と言語



- ✓言語情報は文字情報である
- ✓言語学は文字を扱わない

# システムの構成



# システム実装 + 開発環境

## Ajax ( JavaScript + XML ) 技術

- ユーザーはWebブラウザのみ
- インクリメンタルサーチ (非同期通信の利用)
- ダイナミックHTMLで動的にページの一部を書き換え

## 開発環境

Google Web Toolkit (GWT)

## ミドルウェア

Apache 2.0.59

Apache Tomcat 5.5.26

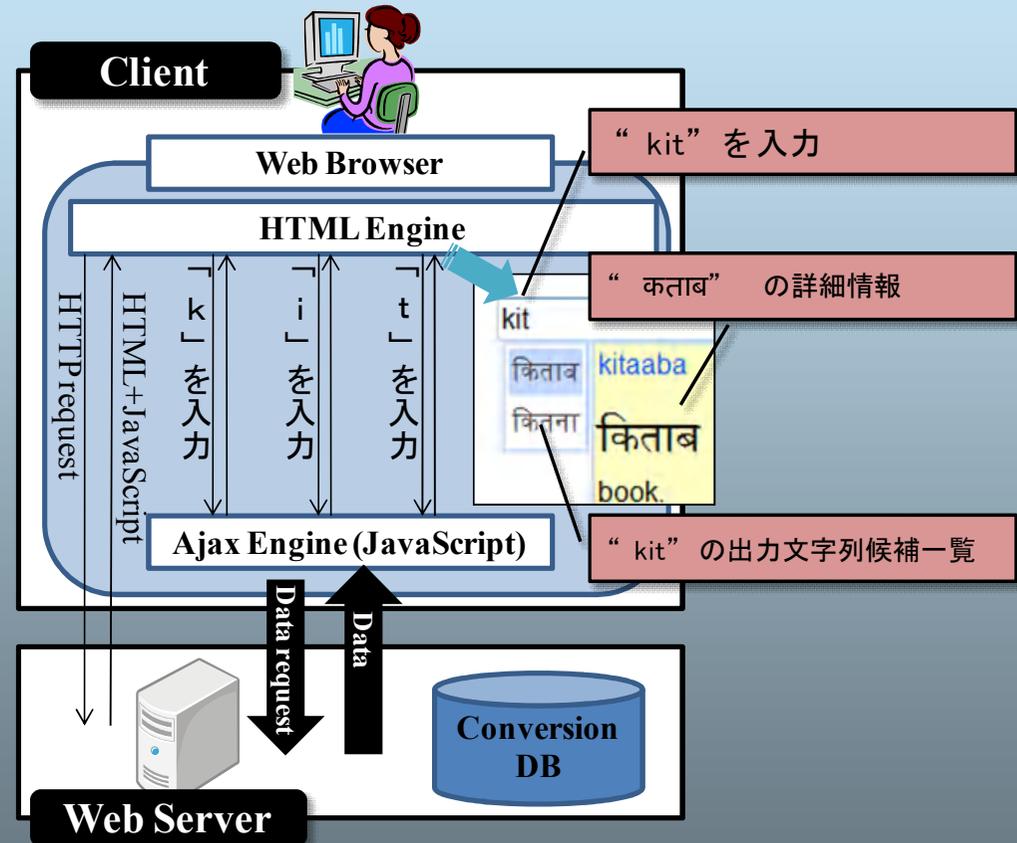
PHP 5.2.2

phpMyAdmin 2.10.1

## DB

MySQL 5.0.82

(変換辞書データを収納)



# メタスクリプト

キーの対応	metascript	目的の文字
一対一(one-to-one)	a	a
一対多(one-to-many)	a	a, <sup>a</sup> a, à, á, â, ã, ä, ā, ɛ, α, ρ, あ, 阿, 亜, अ, ʌ, ʘ, 𑄎, 𑄏, 𑄐
多対一(many-to-one)	kha, k'a, Ka	ख
多対多(many-to-many)	a, A	a, <sup>a</sup> a, à, á, â, ã, ä, ā, ɛ, α, ρ, あ, 阿, 亜, अ, ʌ, ʘ, 𑄎, 𑄏, 𑄐

特徴の対応	metascript	目的の語
アブジャッド(線形抽象化)	trnsln	translation
逆引き(線形方向)	ijom...	西夏文字, 仮名文字, 女文字, 楔形文字, 鏡文字, ...
意味対応(シソーラス、連想)	book	livre, libro, Buch, к н и г а, पु तक, كتاب, 本, 书籍, หนังสือ,

# A. ウイグル語 (アラビア文字) 入力例

The screenshot displays the Ajax IME (Arabic Script) interface. At the top left, there is a logo for 'sorti departu alids' and the title 'Ajax IME (Arabic Script)'. To the right of the title is a globe icon with 'ARABIC LANGUAGE' and 'ARABIC SCRIPT' written on it. Below the title, there are three buttons for font size: 'Small', 'Medium', and 'Large'. A 'resources:' dropdown menu is set to 'Uyghur-Japanese Dic'. There are two checked checkboxes: 'extensive information' and 'Ajax IME mode'. A search input field contains the text 'kitap|'. To the right of the input field is a 'Google Search' dropdown menu, and further right are 'Go' and 'Clear' buttons. A dropdown menu is open below the search input, listing several Uyghur words starting with 'kitap': 'كى تاپ', 'كى تاپباز', 'كى تاپبازلىق', 'كى تاپچە', 'كى تاپچىلىق', 'كى تاپپۇرۇش', 'كى تاپپۇرۇشلۇق', 'كى تاپخان', 'كى تاپخانا', and 'كى تاپخانلىق'. The first item, 'كى تاپ', is highlighted in blue. To the right of this item, the word 'kitap' is shown in blue, and below it, the Japanese character '本' (hon) is shown in a yellow box.

# A. パンジャービー語 (グルムキー文字)入力例

 **Ajax IME (Indic Scripts)** 

font size

resources:  

extensive information  
 Ajax IME mode

ਪੰਜ	paMjaa
ਪੰਜ ਕੱਕੇ	ਪੰਜਾ ਪੰਜਾ <sup>1</sup> /pājā/ [Pers.pāñja] <i>m.</i> {pAj=aa} ❶【身体】手足の五本の指の集まり, 手, てのひら, 掌, たなごころ。❷【身体】手形, てのひらの跡。❸【身体】(足または靴の)爪先の部分。❹【身体】【動物】(動物の)前足。❺【身体】【動物】(猫, ワシ, タカなどの)つめ。❻【比喩】掌握。❼【比喩】支配。
ਪੰਜ ਨਦ	
ਪੰਜ-ਭੁਜ	
ਪੰਜ-ਮੁਖੀ	
ਪੰਜ-ਸਾਲਾ	ਪੰਜਾ ਪੰਜਾ <sup>2</sup> /pājā/ [Pers.pāñja] <i>ca.num.(m.)</i> {pAj=aa} ▶ ਪੰਜਾ 数字の5。
ਪੰਜ-ਸੇਰੀ	
ਪੰਜ-ਤਾਰਾ	
ਪੰਜਾ	
ਪੰਜਾਬ	

# A. 逆引き入力

 **Ajax IME (SSD Japanese)** 

font size

resources:

extensive information  
 Ajax IME mode

<input type="text" value="iakes"/>	iakes
<04> 【世界】	<00>+せ_かい
<04> 【▼娑婆世界】	<04>【世界】
<04> 【《田舎》世界】	<10> ∈ 1 ⇒
<04> 【器世界】	<10>地球上のすべての国家・すべての地域。全人類社会。
<04> 【金色世界】	<11>「—の平和」
<04> 【浄▼瑠▼璃世界】	<11>「—最高の山」
<04> 【東方浄▼瑠▼璃世界】	<10> ∈ 2 ⇒
<04> 【第三世界】	<10>物体や生物など実在する一切のものを含んだ無限の空間。宇宙。哲学では社会的精神的事象をも含める。また、思考・認識する自我に対する客観的世界をさすことも多い。
<04> 【無辺世界】	<11>「可能—」
<04> 【大千世界】	<11>「—の創造」
<04> 【三千大千世界】	<10> ∈ 3 ⇒
<04> 【小千世界】	<10>自分を中心とした生活の場。自分の知識・見聞の範囲。生活圏。世の中。
<04> 【中千世界】	<11>「新し—が開ける」

# Unicodeにおける合成文字の例

1)					$\bar{a}$ U+01DF	→	$\bar{a}$
2)			$\ddot{a}$ U+00E4	+	$\bar{\circ}$ U+0304	→	$\bar{a}$
3)	$a$ U+0061	+	$\ddot{\circ}$ U+0308	+	$\bar{\circ}$ U+0304	→	$\bar{a}$

# A. 「が」のUnicode 正規化表現

 **Code Detection** 

font size

---

Select one of normalization forms.

your input character	NFC	NFD	NFKC	NFKD	normalization form
が U+304C	が U+304C	か+ ◻ U+304B U+3099	が U+304C	か+ ◻ U+304B U+3099	no change <input type="button" value="v"/>

# A. ヒンディー語のつづりチェック例

font size

resources:

Retrieval Options

Except for the initial position  Whole String

**Result:**

hit word	distance	select
पुस्तिका	1	<input checked="" type="radio"/>
पुस्तक	1	<input type="radio"/>
पुस्तकी	1	<input type="radio"/>

# 拡張システム

A. 入力装置



B. 検索フェーズ

全文検索のオプション

- ✓通常文字列
- ✓正規表現
- ✓シソーラス
- ✓共起制限

C. 言語解析ツール

バッチ処理

- ✓タグ付け
- ✓置換(正規表現含む)

# B. 「blue」→「青」

## Search Options

Basic

Regex

Word-set

Co-occurrence

## Enter the item

## Output Options

Hit Count

Results to be Displayed

Hit String

## Result

Your search located 67 occurrences

No.1) Title:Sinhala\_090707

indraniila maaNikyaya ඉන්ද්‍රනිල මාණික්‍යය

[indraniila māṇikyaya][名]《鉞》サファイア, 青玉(類) නිල කැටය

No.2) Title:Sinhala\_090707

# C. 多言語・多文字の簡易辞書

The screenshot shows the CoCo Multiple Processing (ssd) web interface. At the top, there are logos for 'sorti departu alide' and 'MAGADE LAN'. Below the title, there are font size options (Small, Medium, Large) and a resources dropdown menu set to 'ssd 10 languages (with no space)'. A checkbox labeled 'I have an input file.' is present. The main area is divided into two text boxes. The left box contains the word 'ภาษา' and its translations: 'lengua', 'язык', 'linguagem idioma', '언어', 'lingua', 'langue', 'Sprache', and '语言'. Below this, a line of text shows the word 'ภาษา' followed by all the translations: 'ภาษา lengua язык idioma 언어 lingua langue Sprache 语言'. A 'Clear' button is at the bottom left. The right box contains the word '言葉 (言語) 【タイ】' and its translations: '言語【西語】', '言語【露語】', '言葉 (言語) 【葡語】 言語【葡語】', '言語【韓国】', '言語【伊語】', '言語【仏語】', '言語【独語】', and '言語【中国】'. Below this, a line of text shows the word '言葉 (言語) 【タイ】' followed by all the translations: '言葉 (言語) 【タイ】 言語【西語】 言語【露語】 言語【葡語】 言語【韓国】 言語【伊語】 言語【仏語】 言語【独語】 言語【中国】'. A 'GO' button is between the two boxes, and a 'Clear' button is at the bottom right.

# C. 解析例



## CoCo Multiple Processing (Hindi)



font size

Small

Medium

Large

resources:

Hindi2010/03/26

I have an input file.

वह यहाँ आ रहा होगा।

वह यहाँ नहीं आ रहा होगा।

```
+{pron}वह{pron}={0}+ +{adv}  
यहाँ{adv}={0}+ +{iv_stem}आ  
{iv_stem}=[cont_fut]{?_m_sg}  
{punc}|{punc}
```



GO

```
+{pron}वह{pron}={0}+ +{adv}  
यहाँ{adv}={0}+ +{iv_stem}आ  
{iv_stem}=[NGT][cont_fut]{?  
_m_sg} {punc}|{punc}
```

# 最後に、検索プロセスと文字コード列

