総務省SCOPE成果発表 2010/06/11

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

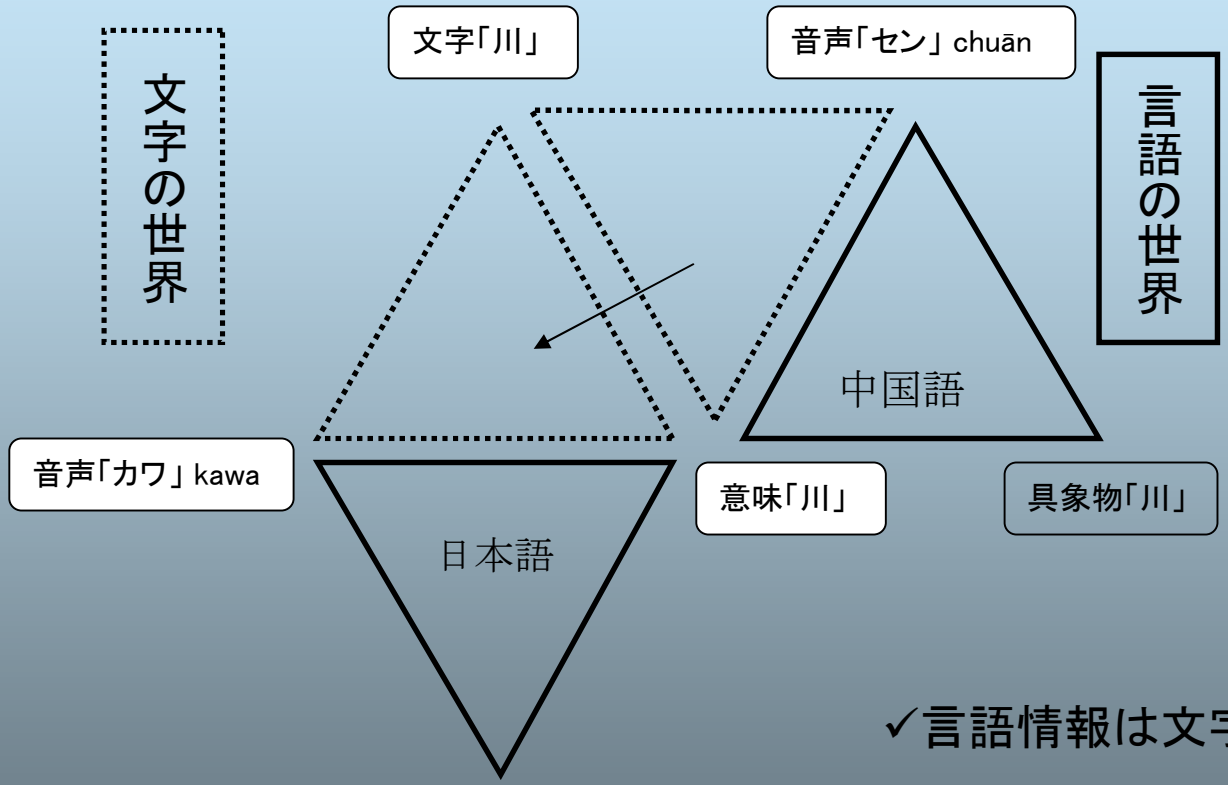
2) 多文字環境に適したInput Method

- 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）

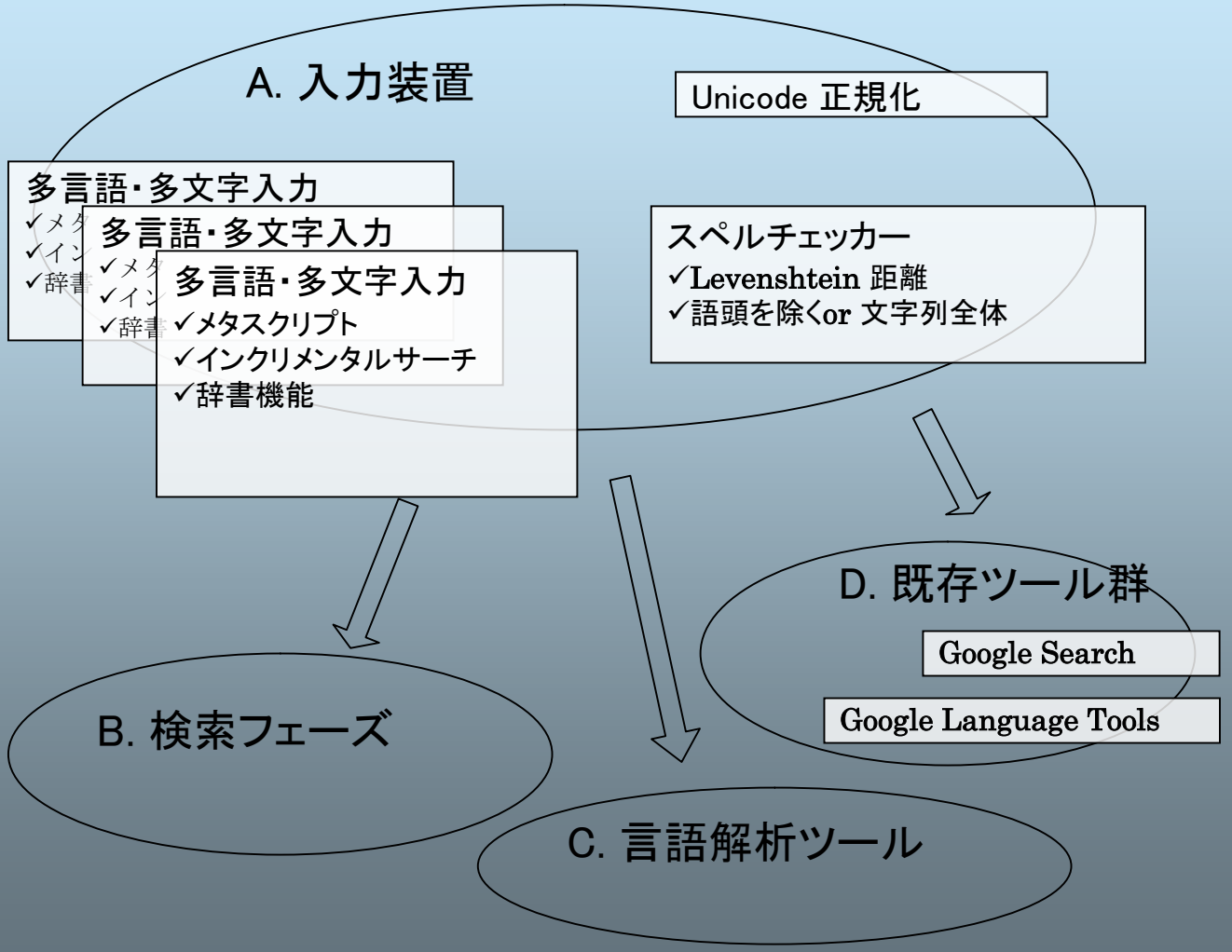
なぜ文字？　文字と言語



✓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


A. 入力装置
Unicode 正規化
多言語・多文字入力
多言語・多文字入力
多言語・多文字入力
✓メタスクリプト
✓インクリメンタルサーチ
✓辞書機能
スペルチェッカー
✓Levenshtein 距離
✓語頭を除くor 文字列全体
D. 既存ツール群
Google Search
Google Language Tools
B. 検索フェーズ
C. 言語解析ツール

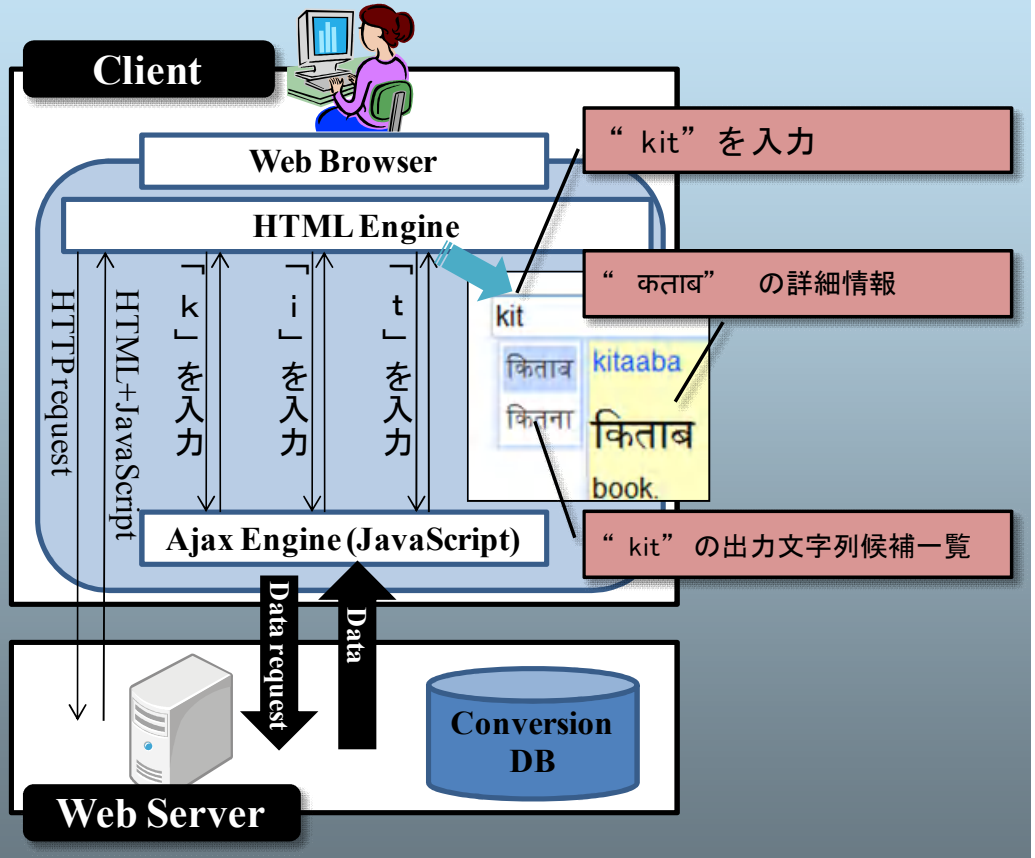
システム実装＋開発環境

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
Apache2.0.59
Apache Tomcat 5.5.26
PHP5.2.2
phpMyAdmin2.10.1

DB
MySQL 5.0.82
（変換辞書データを収納）

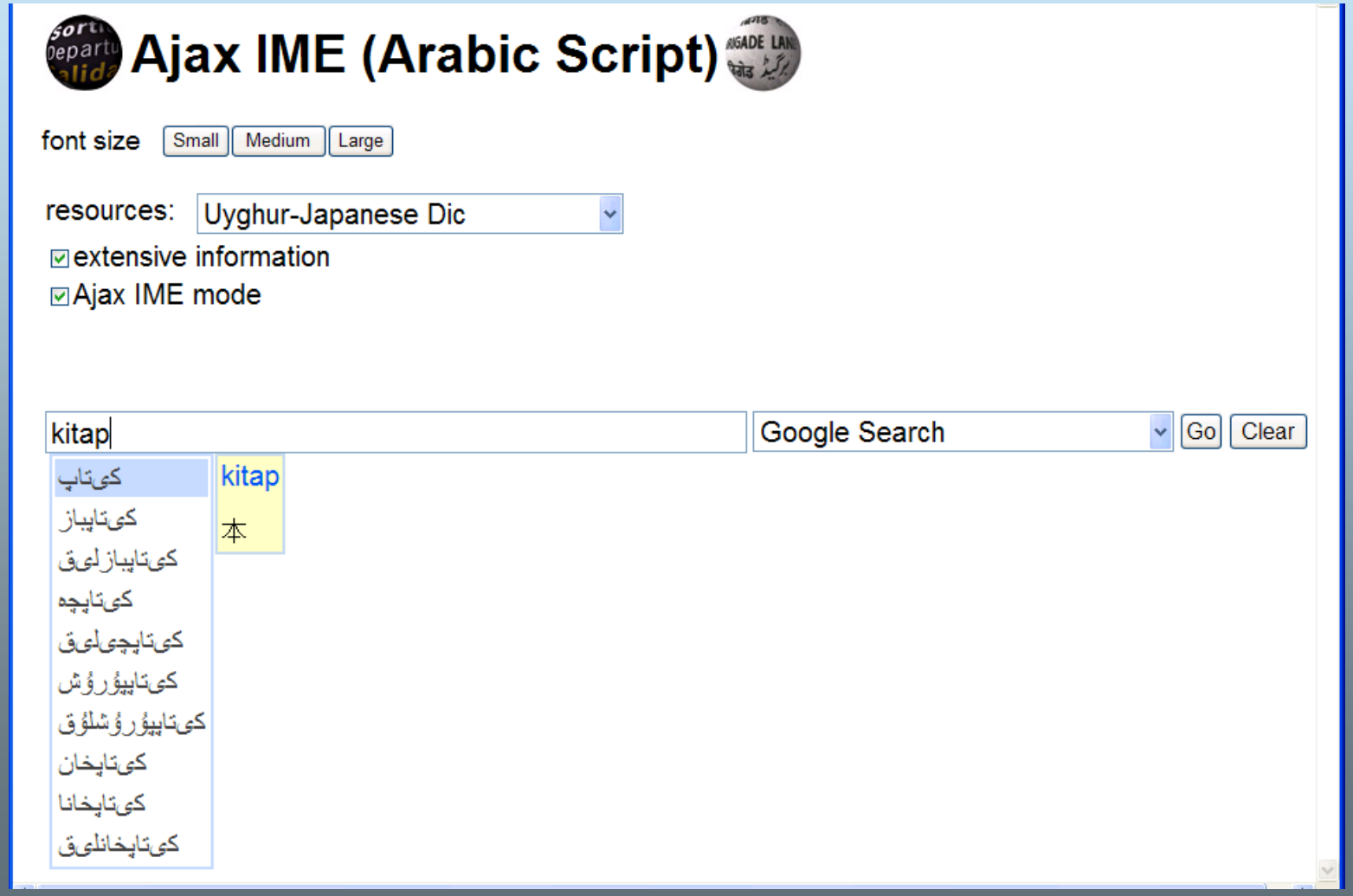


メタスクリプト

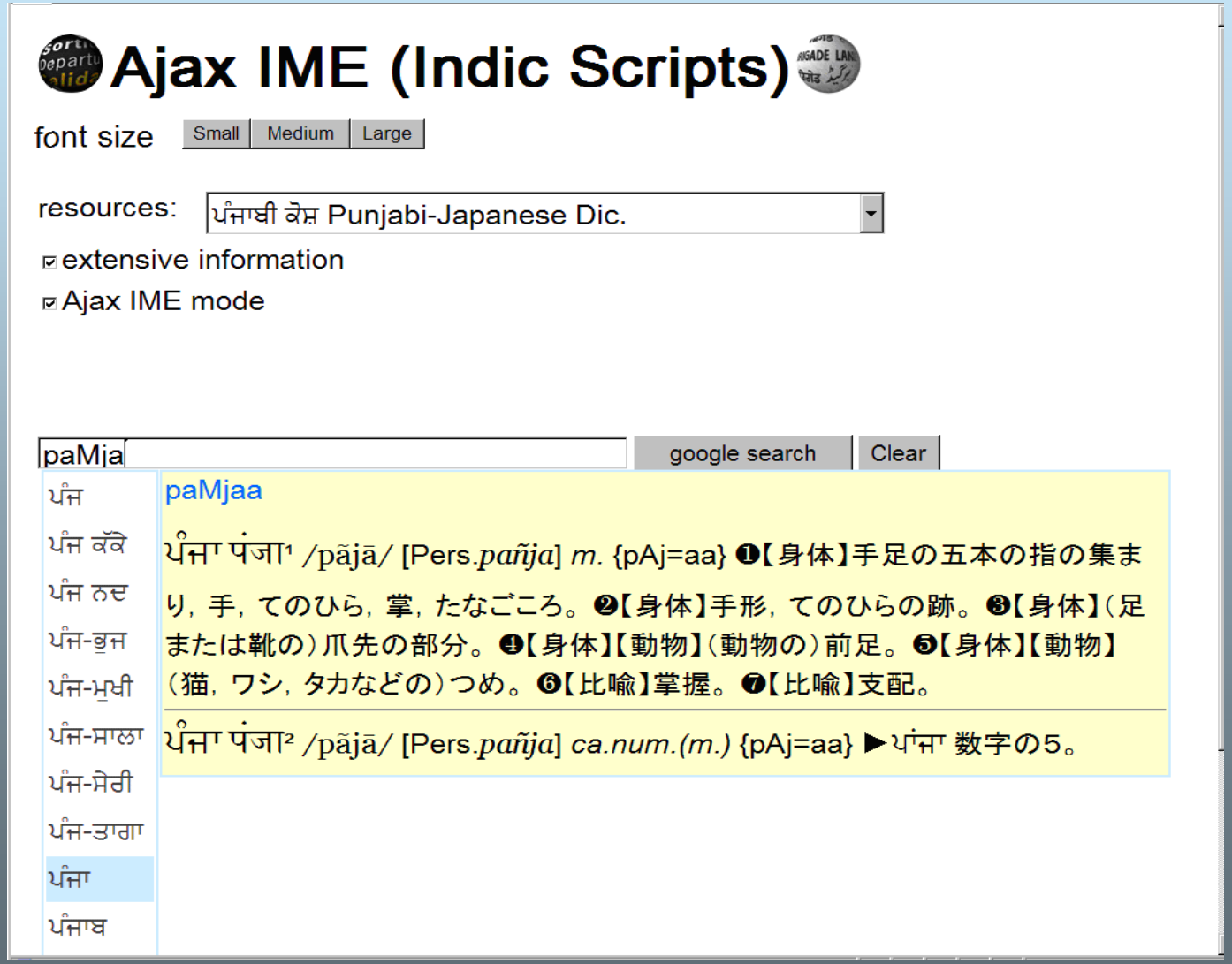
キーの対応	metascript	目的の文字
一対一(one-to-one)	a	a
一対多(one-to-many)	a	a, ª, à, á, â, ã, ä, ā, ɐ, ɑ, ɒ, あ, 阿, 亜, अ, অ, ਅ, அ, ಅ, ๅ
多対一(many-to-one)	kha, k'a, Ka	ख
多対多(many-to-many)	a, A	a, ª, à, á, â, ã, ä, ā, ɐ, ɑ, ɒ, あ, 阿, 亜, अ, অ, ਅ, அ, ಅ, ๅ

特徴の対応	metascript	目的の語
アブジャッド(線形抽象化)	trnsltn	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



A. パンジャービー語（グルムキー文字）入力例



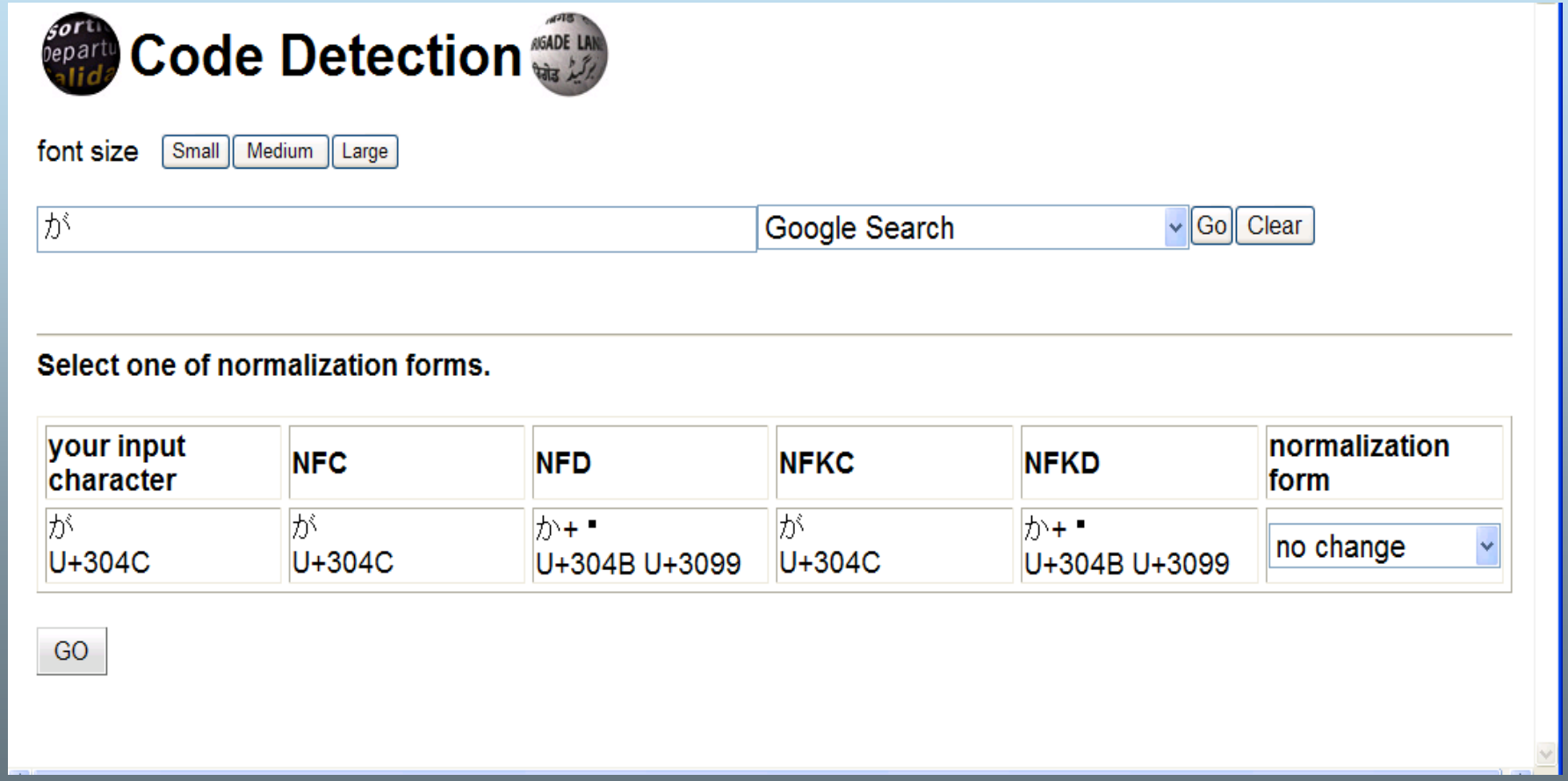
A. 逆引き入力


Ajax IME (SSD Japanese)
font size
Small
Medium
Large
resources:
jirin_r
extensive information
Ajax IME mode
iakes
google search
Clear
<04>【世界】
<04>【▾娑婆世界】
<04>【《田舎》世界】
<04>【器世界】
<04>【金色世界】
<04>【浄▾瑠▾璃世界】
<04>【東方浄▾瑠▾璃世界】
<04>【第三世界】
<04>【無辺世界】
<04>【大千世界】
<04>【三千大千世界】
<04>【小千世界】
<04>【中千世界】
iakes
<00>†せ_かい
<04>【世界】
<10>∈１∋
<10>地球上のすべての国家・すべての地域。全人類社会。
<11>「―の平和」
<11>「―最高の山」
<10>∈２∋
<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。
<11>「可能―」
<11>「―の創造」
<10>∈３∋
<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。

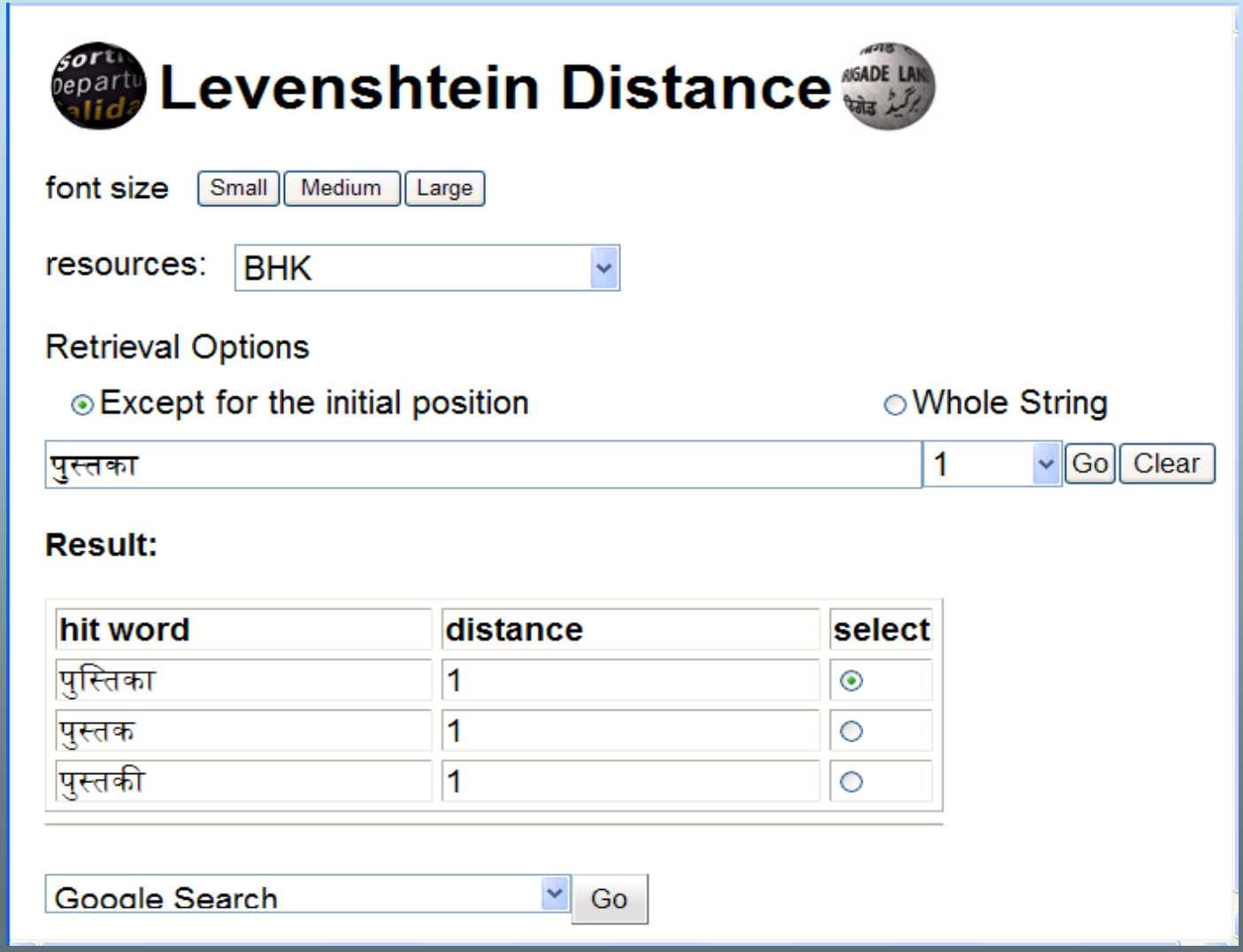
Unicodeにおける合成文字の例

1)					ǟ U+01DF	→	ǟ
2)			ä U+00E4	+	◌̄ U+0304	→	ǟ
3)	a U+0061	+	◌̈ U+0308	+	◌̄ U+0304	→	ǟ

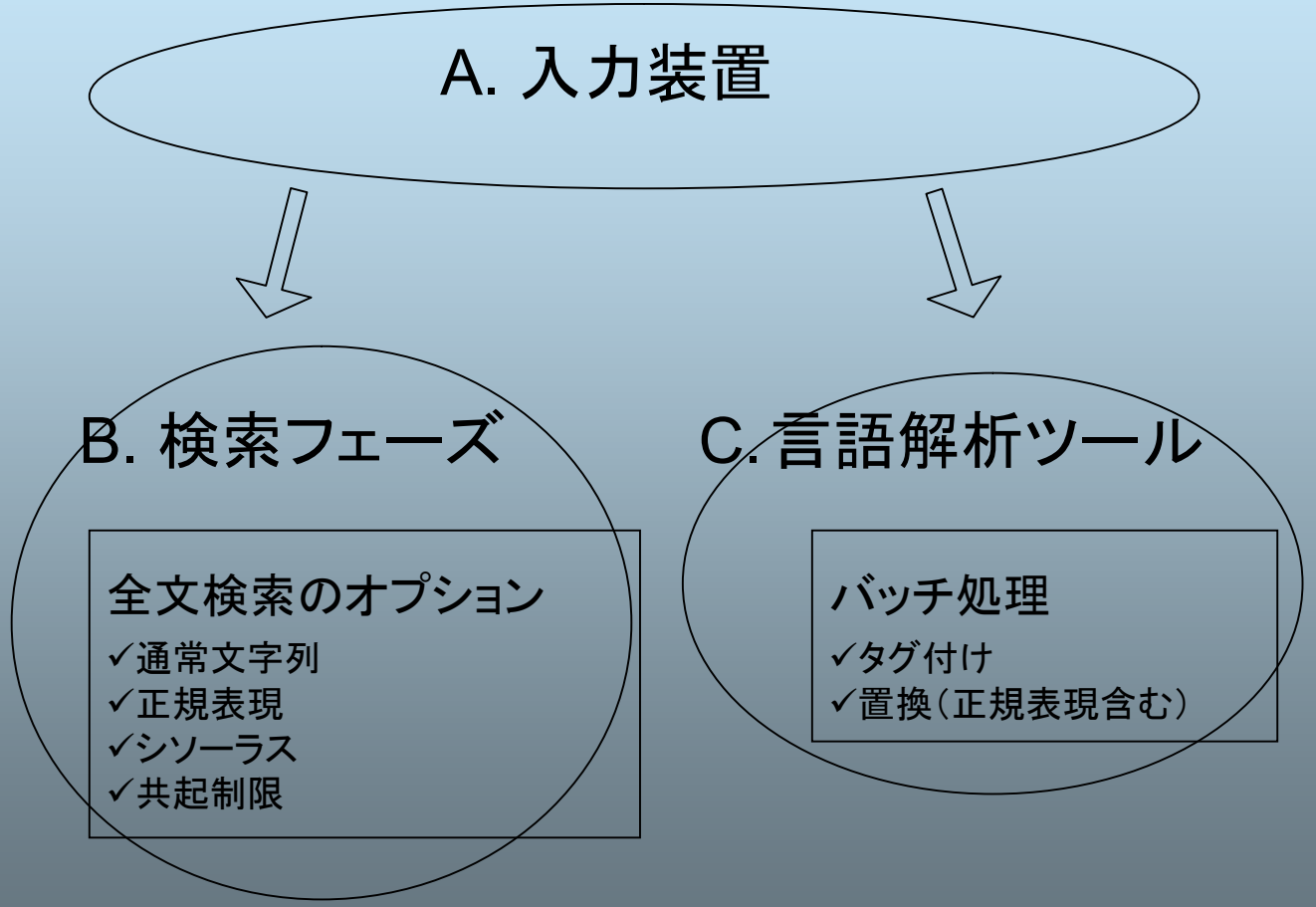
A.「が」のUnicode 正規化表現



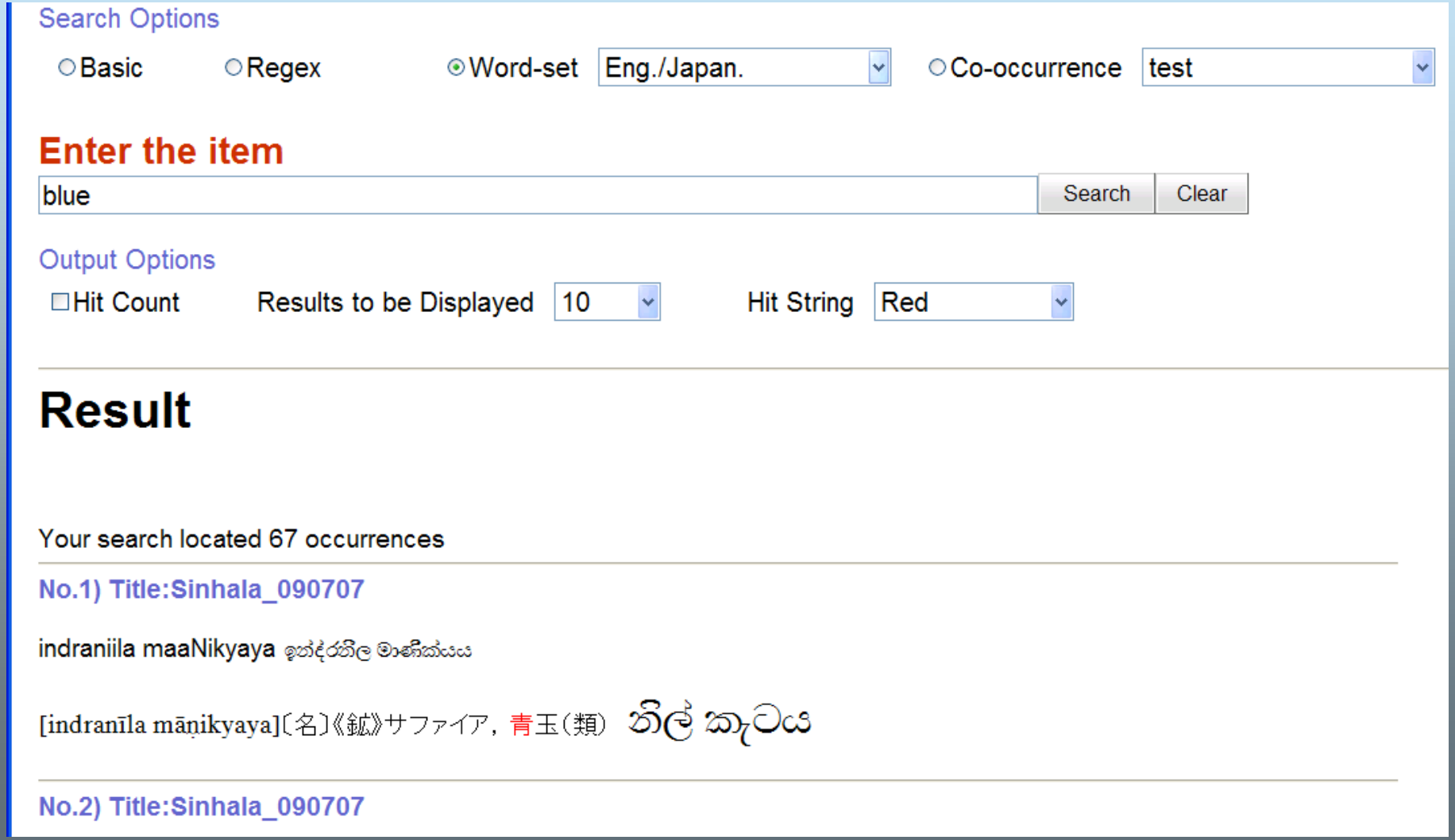
A. ヒンディー語のつづりチェック例



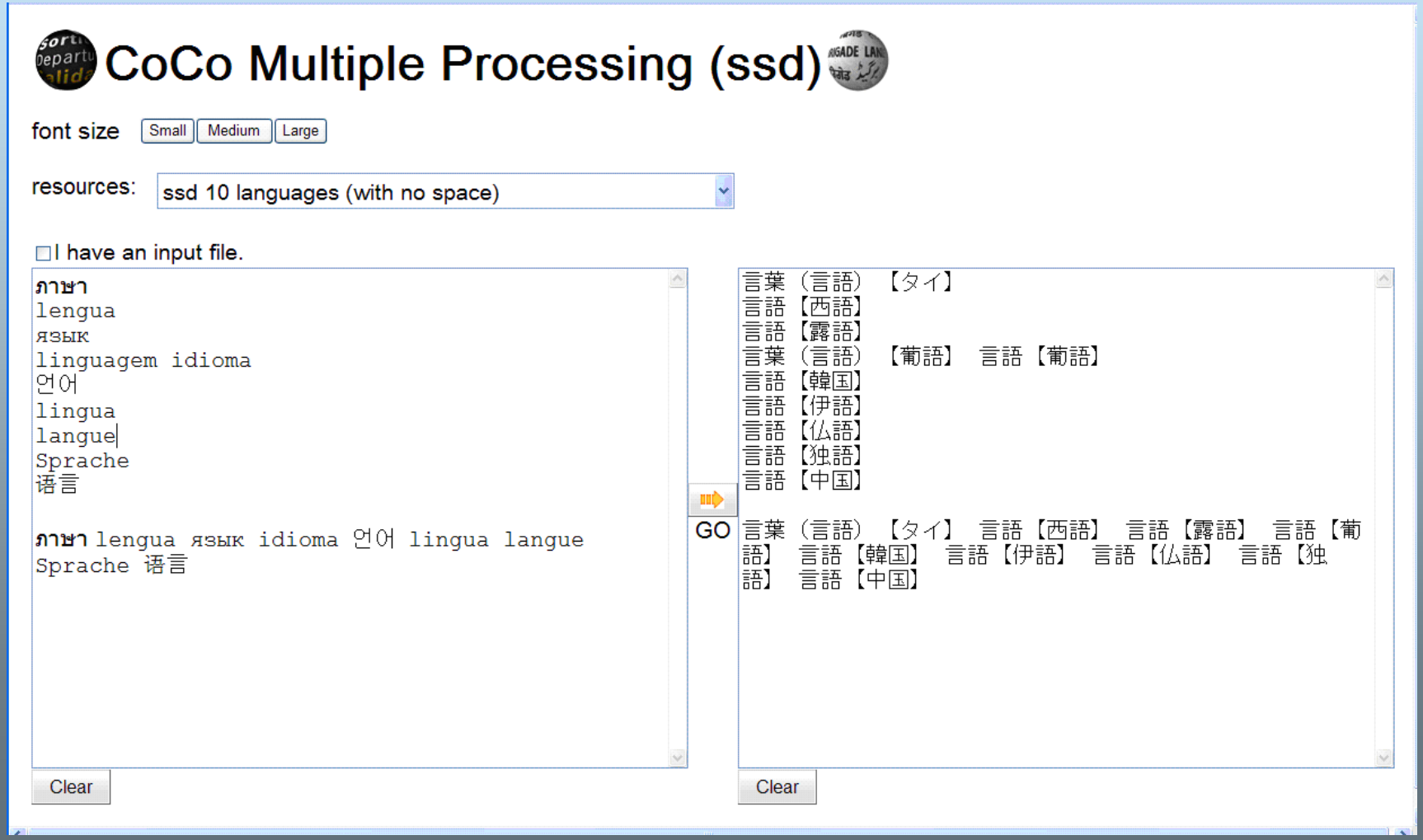
拡張システム


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）

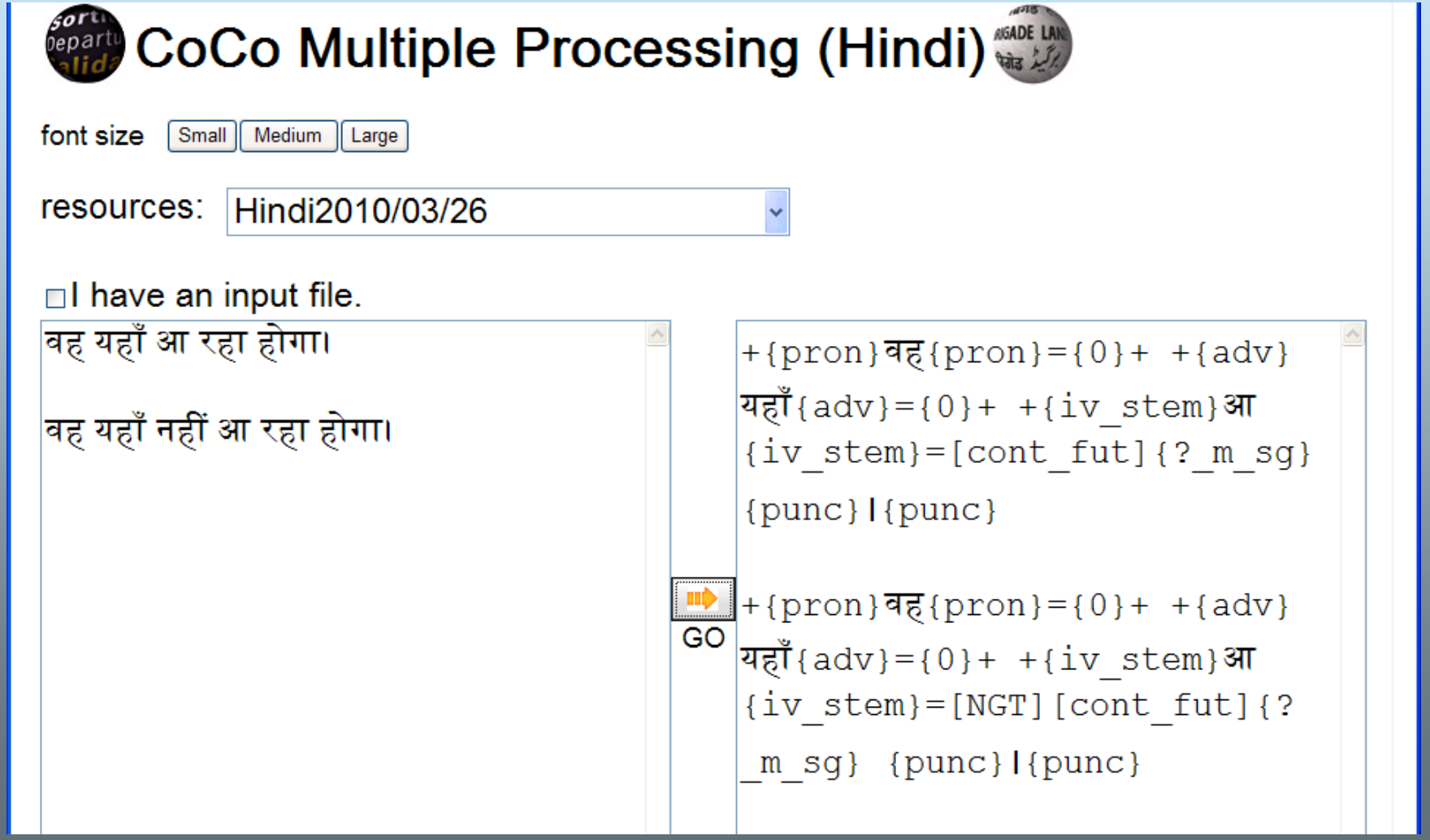
B.「blue」→「青」


Search Options
Basic
Regex
Word-set
Eng./Japan.
Co-occurrence
test
Enter the item
blue
Search
Clear
Output Options
Hit Count
Results to be Displayed
10
Hit String
Red
Result
Your search located 67 occurrences
No.1) Title:Sinhala_090707
indraniila maaNikyaya
[indranīla māṇikyaya]〔名〕《鉱》サファイア，青玉（類）
No.2) Title:Sinhala_090707

C. 多言語･多文字の簡易辞書



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