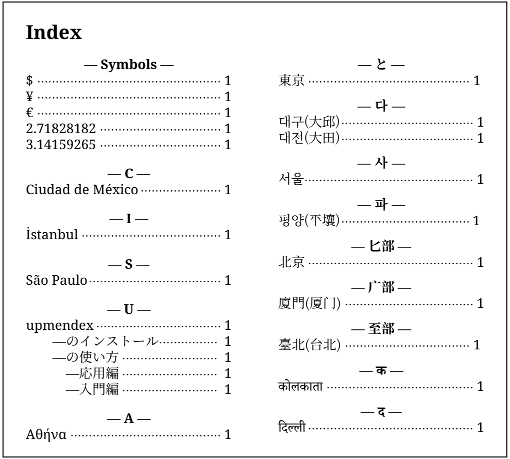
# upmendexで作る多言語索引 Multilingual index processing by upmendex

田中 琢爾
TANAKA Takuji

2025 年 6 月 14 日

# Overview

- Feature of **upmendex**
  — Multilingual index processor
- Localization
  - Latin, Cyrillic, Greek
  - CJK (Chinese, Japanese, Korean)
  - Brahmic (Devanagari, Thai, ···)
  - Arabic, Hebrew
- Multilingual environment
- Benchmark



## Index

— Symbols —

$ 1
¥ 1
€ 1
2.71828182 1
3.14159265 1

— C —

Ciudad de México 1

— I —

İstanbul 1

— S —

São Paulo 1

— U —

upmendex 1
—のインストール 1
—の使い方 1
—応用編 1
—入門編 1

— A —

Αθήνα 1

— と —

東京 1

— 다 —

대구(大邱) 1
대전(大田) 1

— 사 —

서울 1

— 파 —

평양(平壤) 1

— 匕部 —

北京 1

— 广部 —

廈門(厦门) 1

— 至部 —

臺北(台北) 1

— क —

कोलकाता 1

— द —

दिल्ली 1

# Feature of upmendex (Ver. 1.20)

- Index processor
  - Upper compatible with **MakeIndex**/**Mendex**
  - Work with upLaTeX/LuaLaTeX/XeLaTeX
- Localization
  - Support 72 Languages / 22 Scripts
    - Latin (incl. non-English), Cyrillic, Greek
    - CJK (Chinese, Japanese, Korean)
    - Brahmic (Devanagari, Bengali, Thai, ⋯)
    - Arabic, Hebrew
    - Symbol, Number
- Multilingualization
  - Unicode, UTF-8
  - ICU† (Collation, Case Conversion, Category Property)
  - Environment for babel/polyglossia

† ICU: International Components for Unicode

# Language, Script, Locale

| Language | Script / Index | ICU locale |
|---|---|---|
| English | Latin | root |
| Spanish | Latin | es |
| German | Latin | de |
| Turkish | Latin | tr |
| … | | |
| Russian | Cyrillic | ru |
| Ukrainian | Cyrillic | uk |
| … | | |
| Greek | Greek | el |
| Chinese | Hanzi / Pinyin | zh |
| | Hanzi / Stroke | zh-u-co-stroke |
| | Hanzi / Radical-Stroke | zh-u-co-unihan |
| | Hanzi / Zhuyin | zh-u-co-zhuyin |
| Japanese | Kana & Hanzi / Kana | ja |
| Korean | Hangul | ko |

| Language | Script / Index | ICU locale |
|---|---|---|
| Hindi | Devanagari | hi |
| Marathi | Devanagari | mr |
| Bengali | Bengali | bn |
| … | | |
| Thai | Thai | th |
| Persian | Arabic | fa |
| Arabic | Arabic | ar |
| … | | |
| Hebrew | Hebrew | he |
| Yiddish | Hebrew | yi |
| Common | Symbol | |
| | Number | |

**upmendex** supports 72 languages, 22 scripts & 109 locales.

# Latin, Cyrillic, Greek

- Sorting (Collation) | ソート順
- Diacritical mark | ダイアクリティカルマーク
- Digraph/Trigraph | ダイグラフ/トライグラフ

# German, Phonebook Sort Order

## Collation Rule locale: de-u-co-phonebk

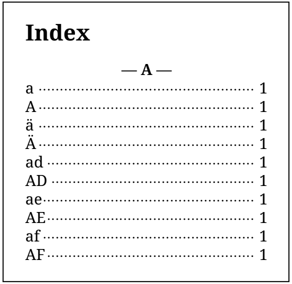
```
&AE<<ä<<<Ä
&OE<<ö<<<Ö
&UE<<ü<<<Ü
. . .
```

## Style File *.ist locale: de-u-co-phonebk

```
icu_locale "de-u-co-phonebk"
```

## German Inputs in *.tex

```
ad\index{ad}.
ae\index{ae}.
AE\index{AE}.
ä\index{ä}.
Ä\index{Ä}.
af\index{af}.
. . .
```

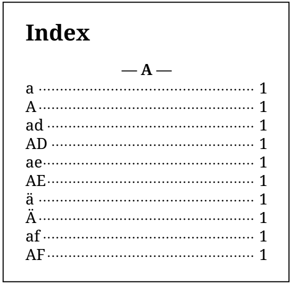


**Index**

— A —

a ........ 1
A ........ 1
ä ........ 1
Ä ........ 1
ad ........ 1
AD ........ 1
ae ........ 1
AE ........ 1
af ........ 1
AF ........ 1

default



**Index**

— A —

a ........ 1
A ........ 1
ad ........ 1
AD ........ 1
ae ........ 1
AE ........ 1
ä ........ 1
Ä ........ 1
af ........ 1
AF ........ 1

phonebook sort order

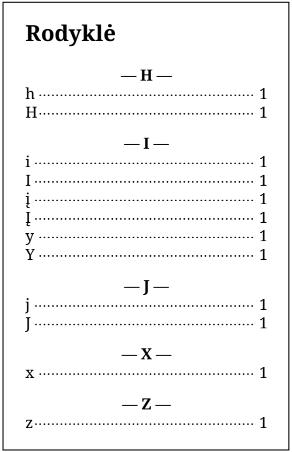
# Lithuanian, Sort Order of Y

## Collation Rule locale: lt

```
&I<<į<<<Į<<y<<<Y
...
```

## Lithuanian Inputs in *.tex

```
i\index{i}.
I\index{I}.
į\index{į}.
Į\index{Į}.
y\index{y}.
Y\index{Y}.
...
```



**Rodyklė**

— **H** —

h ........ 1
H ........ 1

— **I** —

i ........ 1
I ........ 1
į ........ 1
Į ........ 1
y ........ 1
Y ........ 1

— **J** —

j ........ 1
J ........ 1

— **X** —

x ........ 1

— **Z** —

z ........ 1

# Slovak, Diacritical Mark

## Collation Rule locale: sk

```
&O<ô<<<Ô
...
```

## Collation Rule locale: sk-u-co-search

```
&L<ĺ<<<Ĺ<ľ<<<Ľ
&O<ó<<<Ó<ô<<<Ô
...
```

## Slovak Inputs in *.tex

```
l\index{l}.
ĺ\index{ĺ}.
ľ\index{ľ}.
o\index{o}.
ó\index{ó}.
ô\index{ô}.
```

**Index**

— L —

l ........ 1
ĺ ........ 1
ľ ........ 1

— O —

o ........ 1
ó ........ 1

— Ô —

ô ........ 1

default

**Index**

— L —

l ........ 1

— Ĺ —

ĺ ........ 1

— Ľ —

ľ ........ 1

— O —

o ........ 1

— Ó —

ó ........ 1

— Ô —

ô ........ 1

General-Purpose Search

# Turkish, Dotless / Dotted I

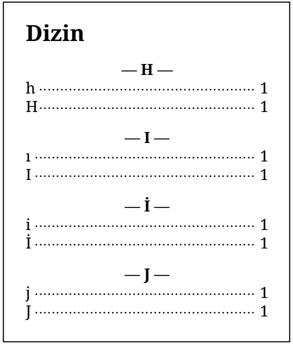
| language | upper | lower |
|---|---|---|
| Turkish | İ | i |
| | I | ı |
| English | I | i |

## Collation Rule locale: tr

```
&[before 1]i<ı<<<I
&i<<<İ
...
```

## Turkish Inputs in *.tex

```
h\index{h}.
H\index{H}.
i\index{i}.
ı\index{ı}.
I\index{I}.
İ\index{İ}.
j\index{j}.
J\index{J}.
...
```



**Dizin**

— **H** —

h ........ 1
H ........ 1

— **I** —

ı ........ 1
I ........ 1

— **İ** —

i ........ 1
İ ........ 1

— **J** —

j ........ 1
J ........ 1

Turkish

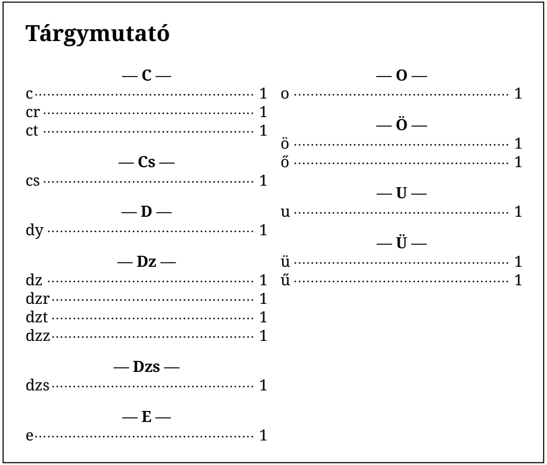
# Hungarian, Digraphs and Trigraph

## Collation Rule locale: hu

```
&C<cs<<<Cs<<<CS
&D<dz<<<Dz<<<DZ
&DZ<dzs<<<Dzs<<<DZS
. . .
```

## Hungarian Inputs in *.tex

```
cr\index{cr}.
cs\index{cs}.
ct\index{ct}.
dy\index{dy}.
dz\index{dz}.
dzr\index{dzr}.
dzs\index{dzs}.
dzt\index{dzt}.
e\index{e}.
. . .
```



**Tárgymutató**

— **C** —

c ........ 1
cr ........ 1
ct ........ 1

— **Cs** —

cs ........ 1

— **D** —

dy ........ 1

— **Dz** —

dz ........ 1
dzr ........ 1
dzt ........ 1
dzz ........ 1

— **Dzs** —

dzs ........ 1

— **E** —

e ........ 1

— **O** —

o ........ 1

— **Ö** —

ö ........ 1
ő ........ 1

— **U** —

u ........ 1

— **Ü** —

ü ........ 1
ű ........ 1

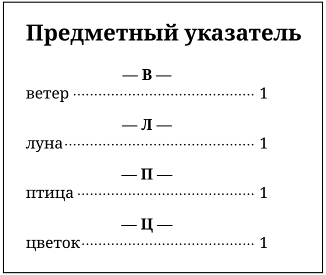
# Cyrillic & Greek

### Russian Inputs in *.tex

```
цветок\index{цветок}.
птица\index{птица}.
ветер\index{ветер}.
луна\index{луна}.
```

### Greek Inputs in *.tex

```
λουλούδι\index{λουλούδι}.
πουλί\index{πουλί}.
άνεμος\index{άνεμος}.
φεγγάρι\index{φεγγάρι}.
```



**Предметный указатель**

— В —

ветер ........ 1

— Л —

луна ........ 1

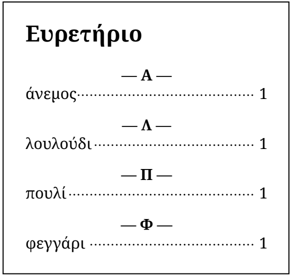
— П —

птица ........ 1

— Ц —

цветок ........ 1

Russian, Cyrillic



**Ευρετήριο**

— Α —

άνεμος ........ 1

— Λ —

λουλούδι ........ 1

— Π —

πουλί ........ 1

— Φ —

φεγγάρι ........ 1

Greek

# CJK (Chinese, Japanese, Korean)

- Chinese: 4 kinds of sort order | 中国語: 4種のソート順
- Japanese: Reading, Extended Kana | 日本語: 読み、仮名拡張
- Korean: composed/decomposed | 韓国語: 完成型・組合型

# Chinese, Han Ideograph Sort Order

## Style File *.ist locale: zh-u-co-unihan

```
%icu_locale "zh"
%icu_locale "zh-u-co-stroke"
icu_locale "zh-u-co-unihan"
%icu_locale "zh-u-co-zhuyin"
hanzi_head  "一部;丨部;丶部;丿部;乙部;亅部;二部;亠部;
...
```

## Chinese Inputs in *.tex

```
花\index{花 (8, 艸, huā, ㄏㄨㄚ)}
鳥\index{鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ)}
風\index{風 (9, 風, fēng, ㄈㄥ)}
月\index{月 (4, 月, yuè, ㄩㄝˋ)}
```

| sort order | | locale |
|---|---|---|
| Pinyin | 拼音 | zh |
| Stroke | 筆畫數 | zh-u-co-stroke |
| Radical-Stroke | 部首筆畫數 | zh-u-co-unihan |
| Zhuyin (Bopomofo) | 注音符號 | zh-u-co-zhuyin |

# Chinese, Han Ideograph Sort Order

## Pinyin Sort Order 拼音 locale: zh

```
\centerline{\bfseries --- F ---}\par\nobreak
  \item 風 (9, 風, fēng, ㄈㄥ)\leaders\hbox{ }\hfill 1

  \indexspace

\centerline{\bfseries --- H ---}\par\nobreak
  \item 花 (8, 艸, huā, ㄏㄨㄚ)\leaders\hbox{ }\hfill 1
```

## Radical-Stroke Sort Order 部首筆畫數 zh-u-co-unihan

```
\centerline{\bfseries --- 月部 ---}\par\nobreak
  \item 月 (4, 月, yuè, ㄩㄝˋ)\leaders\hbox{ }\hfill 1

  \indexspace

\centerline{\bfseries --- 艸部 ---}\par\nobreak
  \item 花 (8, 艸, huā, ㄏㄨㄚ)\leaders\hbox{ }\hfill 1
```

## Stroke Sort Order 筆畫數 zh-u-co-stroke

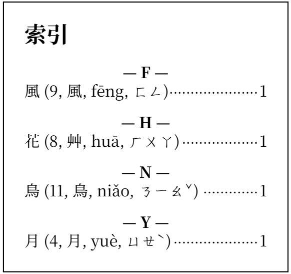
```
\centerline{\bfseries --- 四畫 ---}\par\nobreak
  \item 月 (4, 月, yuè, ㄩㄝˋ)\leaders\hbox{ }\hfill 1

  \indexspace

\centerline{\bfseries --- 八畫 ---}\par\nobreak
  \item 花 (8, 艸, huā, ㄏㄨㄚ)\leaders\hbox{ }\hfill 1
```

## Zhuyin (Bopomofo) Sort Order 注音符號 zh-u-co-zhuyin

```
\centerline{\bfseries --- ㄈ ---}\par\nobreak
  \item 風 (9, 風, fēng, ㄈㄥ)\leaders\hbox{ }\hfill 1

  \indexspace

\centerline{\bfseries --- ㄋ ---}\par\nobreak
  \item 鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ)\leaders\hbox{ }\hfill
```

**upmendex** Output *.ind

# Chinese, Han Ideograph Sort Order



**索引**

**— F —**

風 (9, 風, fēng, ㄈㄥ)………………1

**— H —**

花 (8, 艸, huā, ㄏㄨㄚ)……………1

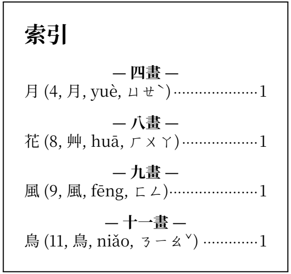
**— N —**

鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ) …………1

**— Y —**

月 (4, 月, yuè, ㄩㄝˋ)………………1

拼音
pinyin



**索引**

**— 四畫 —**

月 (4, 月, yuè, ㄩㄝˋ)………………1

**— 八畫 —**

花 (8, 艸, huā, ㄏㄨㄚ)……………1

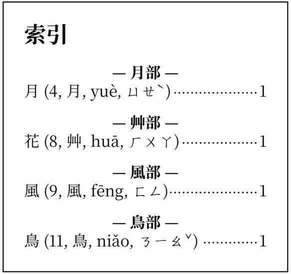
**— 九畫 —**

風 (9, 風, fēng, ㄈㄥ)………………1

**— 十一畫 —**

鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ) …………1

筆畫數
stroke



**索引**

**— 月部 —**

月 (4, 月, yuè, ㄩㄝˋ)………………1

**— 艸部 —**

花 (8, 艸, huā, ㄏㄨㄚ)……………1

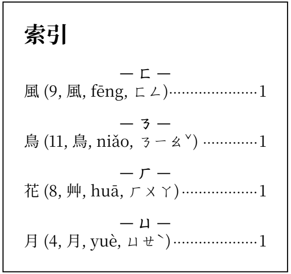
**— 風部 —**

風 (9, 風, fēng, ㄈㄥ)………………1

**— 鳥部 —**

鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ) …………1

部首筆畫數
radical-stroke



**索引**

**— ㄈ —**

風 (9, 風, fēng, ㄈㄥ)………………1

**— ㄋ —**

鳥 (11, 鳥, niǎo, ㄋㄧㄠˇ) …………1

**— ㄏ —**

花 (8, 艸, huā, ㄏㄨㄚ)……………1

**— ㄩ —**

月 (4, 月, yuè, ㄩㄝˋ)………………1

注音符號
zhuyin (bopomofo)

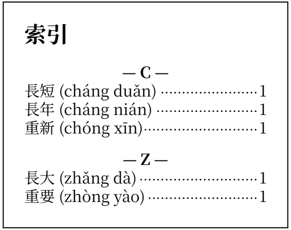
# Chinese, Polyphone (多音字)

## Input with Polyphone in *.tex

```
重新\index{重新 (chóng xīn)}
重要\index{重要 (zhòng yào)}
長年\index{長年 (cháng nián)}
長短\index{長短 (cháng duǎn)}
長大\index{長大 (zhǎng dà)}
```

## Style File *.ist

```
icu_locale "zh"
%icu_locale "zh-u-co-zhuyin"
icu_rules  "&常<<長年<<長短 &崇<<重新"
```



**索引**

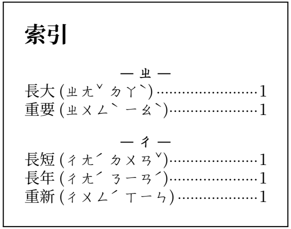
— C —

長短 (cháng duǎn) .................... 1
長年 (cháng nián) .................... 1
重新 (chóng xīn) .................... 1

— Z —

長大 (zhǎng dà) .................... 1
重要 (zhòng yào) .................... 1

拼音
pinyin



**索引**

— ㄓ —

長大 (ㄓㄤˇ ㄉㄚˋ) .................... 1
重要 (ㄓㄨㄥˋ ㄧㄠˋ) .................... 1

— ㄔ —

長短 (ㄔㄤˊ ㄉㄨㄢˇ) .................... 1
長年 (ㄔㄤˊ ㄋㄧㄢˊ) .................... 1
重新 (ㄔㄨㄥˊ ㄒㄧㄣ) .................... 1

注音符號
zhuyin (bopomofo)

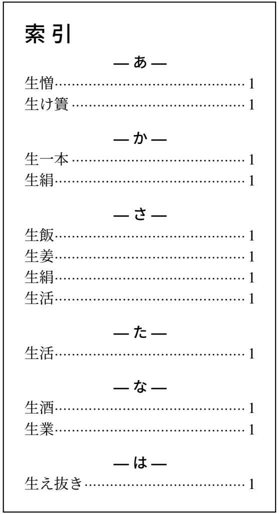
# Japanese, Sort by Reading (Yomi)

## Japanese Inputs with Reading (Yomi) in *.tex

```
\newcommand{\YomiTag}[1]{\relax}

%%  \index{reading@index_word}
生酒\index{なまざけ@生酒}。
生一本\index{きいっぽん@生一本}。
生け簀\index{いけす@生け簀}。
生絹\index{きぎぬ@生絹\YomiTag{きぎぬ}}%
    \index{すずし@生絹\YomiTag{すずし}}。
生飯\index{さば@生飯}。
生姜\index{しょうが@生姜}。
生活\index{せいかつ@生活\YomiTag{せいかつ}}
    \index{たつき@生活\YomiTag{たつき}}。
...
```

Japanese words consist of Hanzi (ideographs) & Kana (syllabaries), sorted by reading (yomi), indexed by Kana.

This feature is implemented by **ASCII mendex**.



**索 引**

— あ —

生憎 ........ 1
生け簀 ........ 1

— か —

生一本 ........ 1
生絹 ........ 1

— さ —

生飯 ........ 1
生姜 ........ 1
生絹 ........ 1
生活 ........ 1

— た —

生活 ........ 1

— な —

生酒 ........ 1
生業 ........ 1

— は —

生え抜き ........ 1

# Japanese, Reading & Dictionary

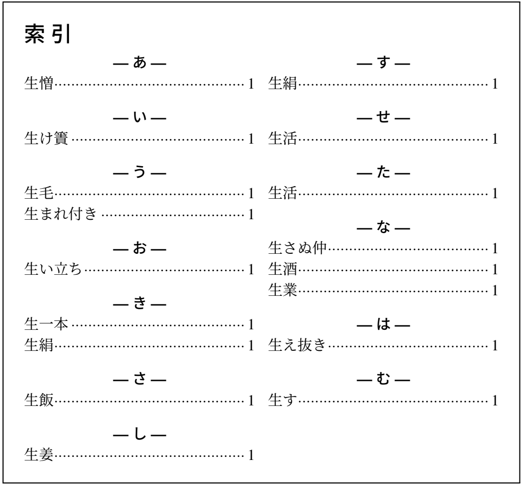
## Japanese Inputs in *.tex

```
生酒\index{生酒}。
生一本\index{生一本}。
生け簀\index{生け簀}。
生絹\index{生絹}%
    \index{すずし@生絹\YomiTag{すずし}}。
生え抜き\index{生え抜き}。
...
```

## Dictionary *.dic

```
index_word  reading
生酒        なまざけ
生一本      きいっぽん
生け簀      いけす
生絹        きぎぬ
生え抜き    はえぬき
...
```

Implemented by **ASCII mendex**.



索 引

— あ —

生憎 ........ 1

— い —

生け簀 ........ 1

— う —

生毛 ........ 1

生まれ付き ........ 1

— お —

生い立ち ........ 1

— き —

生一本 ........ 1

生絹 ........ 1

— さ —

生飯 ........ 1

— し —

生姜 ........ 1

— す —

生絹 ........ 1

— せ —

生活 ........ 1

— た —

生活 ........ 1

— な —

生さぬ仲 ........ 1

生酒 ........ 1

生業 ........ 1

— は —

生え抜き ........ 1

— む —

生す ........ 1

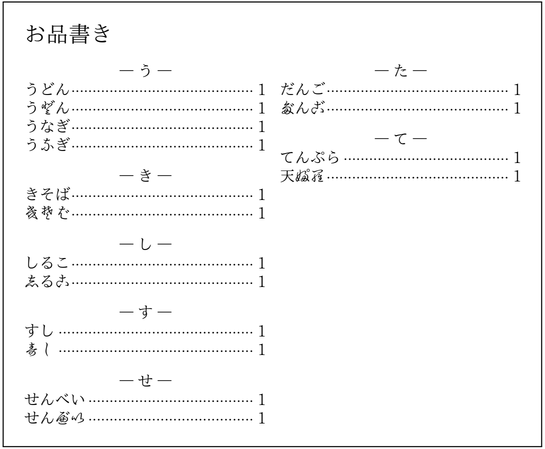
# Japanese, Hentaigana

## Inputs with Hentaigana in *.tex

```
𛀨𛁛𛂞゙\index{𛀨𛁛𛂞゙}
う𛁵ぎ\index{う𛁵ぎ}
𛁆る𛀸\index{𛁆る𛀸}
𛁋𛁄\index{𛁋𛁄}
天𛂰゚𛃭\index{天𛂰゚𛃭}
…
```

## Dictionary for Hentaigana *.dic

| index_word | reading |
|---|---|
| 𛀨 | き |
| 𛁛 | そ |
| 𛂞 | は |
| 𛁵 | な |
| 𛁆 | し |
| 𛀸 | こ |
| … | |



お品書き

— う —

うどん……1
う𛁻゙ん……1
うなぎ……1
う𛁵ぎ……1

— き —

きそば……1
𛀨𛁛𛂞゙……1

— し —

しるこ……1
𛁆る𛀸……1

— す —

すし……1
𛁋𛁄……1

— せ —

せんべい……1
せん𛂫゙𛀆……1

— た —

だんご……1
𛁞゙ん𛀸゙……1

— て —

てんぷら……1
天𛂰゚𛃭……1

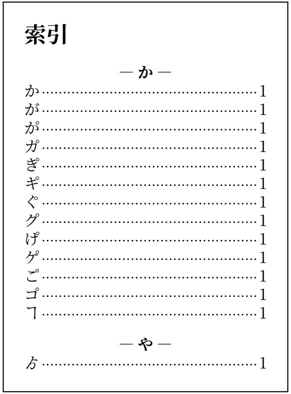
# Extended Kana in JIS X 0213

## Extended Kana Inputs in *.tex

```
か゚\index{か゚}. % NGA
き゚\index{き゚}. % NGI
く゚\index{く゚}. % NGU
...
ゟ\index{ゟ}. % Hiragana Digraph Yori
ヿ\index{ヿ}. % Katakana Digraph Koto
```

## Aynu itak Inputs in *.tex

```
ㇰ\index{ㇰ}
ㇱ\index{ㇱ}
ㇲ\index{ㇲ}
ㇷ゚\index{ㇷ゚}
セ゚\index{セ゚}
ツ゚\index{ツ゚}
ト゚\index{ト゚}
...
```



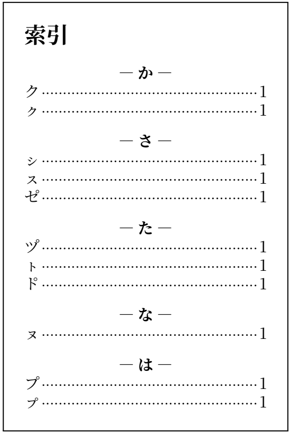
索引

— か —

か……1
が……1
か゚……1
カ゚……1
き゚……1
キ゚……1
く゚……1
ク゚……1
け゚……1
ケ゚……1
こ゚……1
コ゚……1
ヿ……1

— や —

ゟ……1



索引

— か —

ク……1
ㇰ……1

— さ —

ㇱ……1
ㇲ……1
セ゚……1

— た —

ツ゚……1
ㇳ……1
ト゚……1

— な —

ㇴ……1

— は —

プ……1
ㇷ゚……1

# Japanese, Archaic Kana

## Archaic Kana Inputs in *.tex

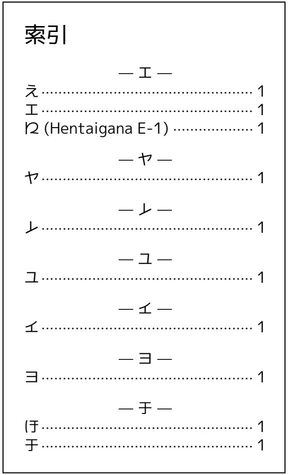
```
𛀁\index{𛀁}. % Hiragana Archaic YE
             %  or Hentaigana E-1
𛄟\index{𛄟}. % Hiragana WU
𛄠\index{𛄠}. % Katakana YI
𛄡\index{𛄡}. % Katakana YE
𛄢\index{𛄢}. % Katakana WU
...
```

## Dictionary: “𛀁” is Hentaigana E-1

𛀁 え

## Style: “𛀁” is Hiragana Archaic YE

```
icu_rules "&ゆ<𛀁<<<𛄡<よ"
```



索引

― エ ―

え .......... 1
エ .......... 1
𛀁 (Hentaigana E-1) .......... 1

― ヤ ―

ヤ .......... 1

― 𛄠 ―

𛄠 .......... 1

― ユ ―

ユ .......... 1

― 𛄡 ―

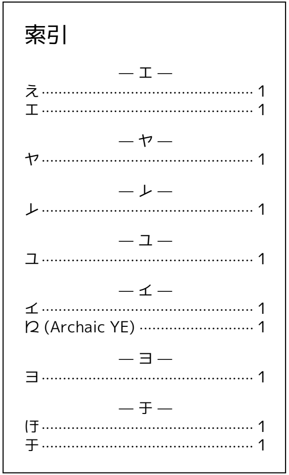
𛄡 .......... 1

― ヨ ―

ヨ .......... 1

― 𛄢 ―

𛄟 .......... 1
𛄢 .......... 1



索引

― エ ―

え .......... 1
エ .......... 1

― ヤ ―

ヤ .......... 1

― 𛄠 ―

𛄠 .......... 1

― ユ ―

ユ .......... 1

― 𛄡 ―

𛄡 .......... 1
𛀁 (Archaic YE) .......... 1

― ヨ ―

ヨ .......... 1

― 𛄢 ―

𛄟 .......... 1
𛄢 .......... 1

# Korean, Modarn / Archaic Hangul

**ideograph**
한자

日

**modarn**
현대 한글

일

U+C77C

**archaic, composed / decomposed**
옛 한글, 완성형 / 조합형 (完成型 / 組合型)

〮ᅀᅵᇙ = ᅀ + ᅵ + ᇙ + 〮

| ᅀ | ᅵ | ᇙ | 〮 |
|---|---|---|---|
| U+1140 | U+1175 | U+11D9 | U+302E |
| 첫소리 | 가운뎃소리 | 끝소리 | 방점 |
| 初聲 | 中聲 | 終聲 | 傍點 |

| | Unicode block | style | | upmendex | upLaTeX | XeLaTeX |
|---|---|---|---|---|---|---|
| modarn | Hangul Syllables | composed | 완성형 | √ | √ | √ |
| ex. 일 | Hangul Jamo | decompopsed | 조합형 | √ | N.A. | √ |
| archaic | Private Use Area | composed | 완성형 | via dictionary | √ | √ |
| ex. 〮ᅀᅵᇙ | Hangul Jamo | decomposed | 조합형 | √ | N.A. | √ |

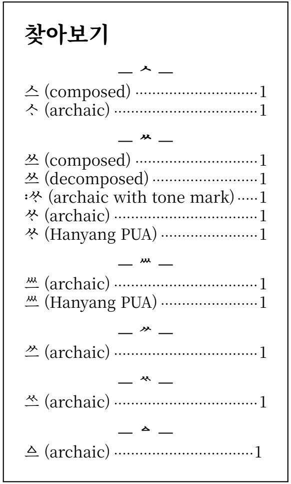
# Korean Hangul

## Hangul Inputs in *.tex

```
쓰\index{쓰 (composed)}.
쓰\index{쓰 (decomposed)}.

ᄊᆞ\index{ᄊᆞ (archaic)}.
ᄊᆞ〯\index{ᄊᆞ〯 (archaic with tone mark)}.
ᄊᆞ\index{ᄊᆞ (Hanyang PUA)}.
...
```

## Dictionary for PUA code *.dic

```
Hanyang PUA code   decomposed
ᄊᆞ                 ᄊᆞ
ᄊᆡ                 ᄊᆞᅵ
...
```



### 찾아보기

— ᄉ —

스 (composed) ........ 1
ᄉᆞ (archaic) ........ 1

— ᄊ —

쓰 (composed) ........ 1
쓰 (decomposed) ........ 1
〯ᄊᆞ (archaic with tone mark) ..... 1
ᄊᆞ (archaic) ........ 1
ᄊᆞ (Hanyang PUA) ........ 1

— ᄴ —

ᄴᅳ (archaic) ........ 1
ᄴᅳ (Hanyang PUA) ........ 1

— ᄽ —

ᄽᅳ (archaic) ........ 1

— ᄿ —

ᄿᅳ (archaic) ........ 1

— ᅀ —

ᅀᅳ (archaic) ........ 1

# Complex Text Layout

- Brahmic scripts:
  Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao
- Arabic, Hebrew: R-to-L typeset
- Symbol, Number

# Devanagari & Thai (experimental)

## Hindi Inputs in *.tex

फूल \index{ फूल }

चिड़िया \index{ चिड़िया }

हवा \index{ हवा }

चांद \index{ चांद }

...

## Thai Inputs in *.tex

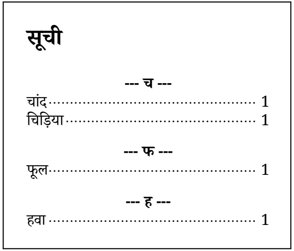
ดอกไม้ \index{ ดอกไม้ }

นก \index{ นก }

ลม \index{ ลม }

ดวงจันทร์ \index{ ดวงจันทร์ }

...



**सूची**

**--- च ---**

चांद ........ 1
चिड़िया ........ 1

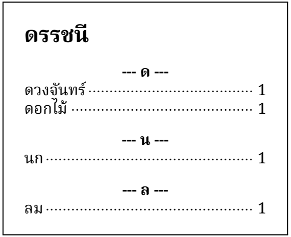
**--- फ ---**

फूल ........ 1

**--- ह ---**

हवा ........ 1

Hindi, Devanagari



**ดรรชนี**

**--- ด ---**

ดวงจันทร์ ........ 1
ดอกไม้ ........ 1

**--- น ---**

นก ........ 1

**--- ล ---**

ลม ........ 1

Thai

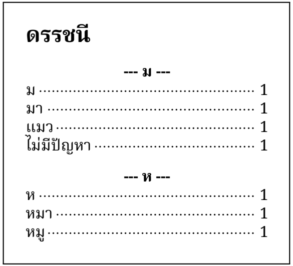
Typeset by XeLaTeX

# Thai/Lao Prevowels Reordering (experimental)

## Thai Inputs in *.tex

```
หมา \index{หมา}
หมู \index{หมู}
แมว \index{แมว}
ไม่มีปัญหา \index{ไม่มีปัญหา}
มา \index{มา}
ห \index{ห}
ม \index{ม}
```

A sequence of a Thai vowel เ แ โ ใ ไ or a Lao vowel ເ ແ ໂ ໃ ໄ followed by a Thai consonant ก ข ค ฆ ง ... or a Lao consonant ກ ຂ ຄ ງ ... is placed after the consonant for collation purposes.



**ดรรชนี**

--- ม ---

ม ........ 1
มา ........ 1
แมว ........ 1
ไม่มีปัญหา ........ 1

--- ห ---

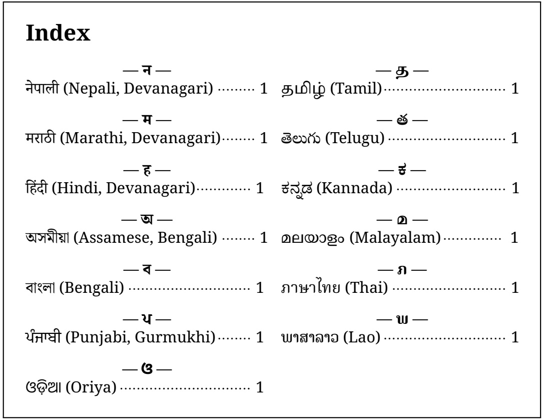
ห ........ 1
หมา ........ 1
หมู ........ 1

# Brahmic Scripts (experimental)

## Multiscript Inputs in *.tex

```
\begin{hindi}
हिंदी \index{ हिंदी (Hindi, Devanagari)}
मराठी \index{ मराठी (Marathi, Devanagari)}
नेपाली \index{ नेपाली (Nepali, Devanagari)}
\end{hindi}
...
\begin{bengali}
বাংলা \index{ বাংলা (Bengali)}
অসমীয়াা \index{ অসমীয়া (Assamese, Bengali)}
\end{bengali}
...
```

Supports 12 Brahmic scripts,
15 languages.



## Index

with XeLaTeX & polyglossia

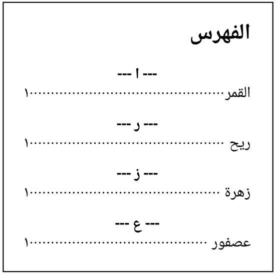
# Arabic & Hebrew (experimental)

## Arabic Inputs in *.tex

زهرة \index{ زهرة }
عصفور \index{ عصفور }
ريح \index{ ريح }
القمر \index{ القمر }
...

## Hebrew Inputs in *.tex

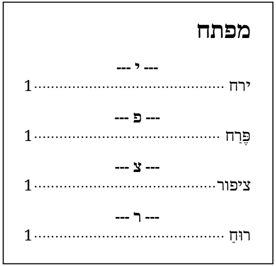
פֶּרַח \index{ פֶּרַח }
ציפור \index{ ציפור }
רוּחַ \index{ רוּחַ }
ירח \index{ ירח }
...



الفهرس

--- ا ---
القمر ........................................ ١
--- ر ---
ريح ........................................ ١
--- ز ---
زهرة ........................................ ١
--- ع ---
عصفور ........................................ ١

Arabic



מפתח

--- י ---
ירח ........................................ 1
--- פ ---
פֶּרַח ........................................ 1
--- צ ---
ציפור ........................................ 1
--- ר ---
רוּחַ ........................................ 1

Hebrew

R-to-L typeset by XeLaTeX.
**upmendex** processes only indexing.

# Symbol, Number

| Script | charType | example | treatment by upmendex |
|---|---|---|---|
| Latin | Lu, Ll, Lo, ... : letters | ABCabcＡ ａⓐ | directly pass to ICU collator |
| Greek | Lu, Ll, Lo, ... : letters | ΑΒΓαβγ | direct |
| Cyrillic | Lu, Ll, Lo, ... : letters | АБВабв | direct |
| Kana | Lo : other letter | あいうアｱ㋐㌀ | direct |
| Hangul | Lo : other letter | 가나다ﾡ㉡㉰㉼ | direct |
| Hanzi | Lo : other letter | 花鳥風月 | lookup dictionary or direct |
| — | Lm : modifier letter | ー゛゜ｰ | direct |
| Number | Nd : dicimal digit number | 012０１２ | direct |
|  | No : other number | ¹²③❹⑤❻🄈⑻⒐ | lookup dictionary or direct |
| Symbol | Sk : modifier symbol | ¨¯´^`゛゜ | direct |
|  | Sm : math symbol | ÷▷♯ | lookup dictionary or direct |
|  | So : other symbol | ☃〠☎♥⚽⚾ | lookup dictionary or direct |
|  | Sc : currency symbol | €$＄¢£¥￥₩￦ | lookup dictionary or direct |
|  | Po, Pd, Mn, Me, ... : other punctuation etc. | ?!!?¡¿†#§¶— | direct |
| — | Cc : control character | ESC, BS, DEL | ignore |
| — | Cf : format character | BOM, RLM | direct |
| others | Lo, etc. (unknown scripts) |  | lookup dic or direct (option “-f”) or ignore |

Characters are classified by Unicode *General Category Values* or “charTypes”

Ref. https://unicode.org/reports/tr44/#General_Category_Values

# Multilingual Environment with upLaTeX/pxbabel

## Block Setting for Scripts in Style File *.ist

```
script_preamble   cyrillic "\n\\fontencoding{T2A}\\selectfont"
script_postamble  cyrillic "\n\\fontencoding{T1}\\selectfont"

script_preamble   hangul   "\n\\begin{otherlanguage}{korean}"
script_postamble  hangul   "\n\\end{otherlanguage}"

script_preamble   hanzi    "\n\\begin{otherlanguage}{tchinese}"
script_postamble  hanzi    "\n\\end{otherlanguage}"
```

## **upmendex** Output *.ind

```
\centerline{\bfseries --- C ---}\par\nobreak
  \item София\leaders\hbox{~}\hfill 1
...
\fontencoding{T1}\selectfont

  \indexspace

\centerline{--- サ ---}\par\nobreak
  \item さいたま\leaders\hbox{~}\hfill {1}
  \item 札幌\leaders\hbox{~}\hfill {1}
...
\begin{otherlanguage}{korean}

  \indexspace

\centerline{\bfseries --- ㅂ ---}\par\nobreak
  \item 부산(釜山)\leaders\hbox{~}\hfill 1
...
\end{otherlanguage}
```

# Multilingual Environment with XeLaTeX/polyglossia

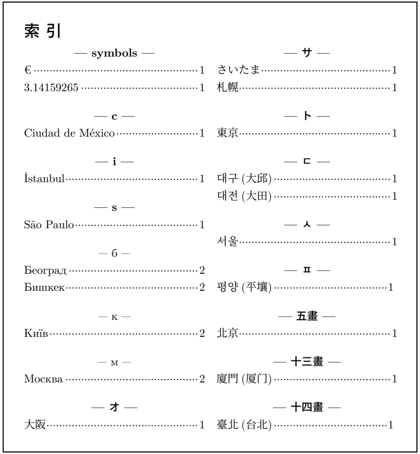
## Block Setting for Scripts in Style File *.ist

```
script_preamble   cyrillic "\n\\begin{russian}"
script_postamble  cyrillic "\n\\end{russian}"

script_preamble   kana     "\n\\begin{japanese}"
script_postamble  kana     "\n\\end{japanese}"

script_preamble   hangul   "\n\n\\begin{korean}"
script_postamble  hangul   "\n\\end{korean}"

script_preamble   hebrew   "\n\\begin{hebrew}"
script_postamble  hebrew   "\n\\end{hebrew}"
```

## **upmendex** Output *.ind

```
\centerline{--- さ ---}\par\nobreak
  \item さいたま\leaders\hbox{~}\hfill {1}
...
\end{japanese}

\begin{korean}

  \indexspace

\centerline{--- ㄷ ---}\par\nobreak
  \item 대구(大邱)\leaders\hbox{~}\hfill {1}
...
\end{korean}
\begin{hebrew}

  \indexspace

\centerline{--- א ---}\par\nobreak
  \item אשדוד\leaders\hbox{~}\hfill {2}
...
\end{hebrew}
```

# Output of Multilingual Index



## 索 引

— symbols —

€ ........ 1

3.14159265 ........ 1

— c —

Ciudad de México ........ 1

— i —

İstanbul ........ 1

— s —

São Paulo ........ 1

— б —

Београд ........ 2

Бишкек ........ 2

— к —

Київ ........ 2

— м —

Москва ........ 2

— オ —

大阪 ........ 1

— サ —

さいたま ........ 1

札幌 ........ 1

— ト —

東京 ........ 1

— ㄷ —

대구 (大邱) ........ 1

대전 (大田) ........ 1

— ㅅ —

서울 ........ 1

— ㅍ —

평양 (平壤) ........ 1

— 五畫 —

北京 ........ 1

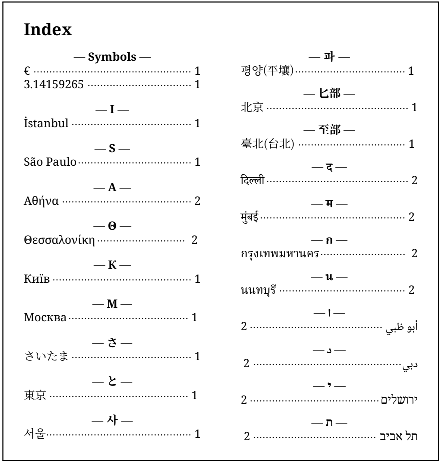
— 十三畫 —

廈門 (厦门) ........ 1

— 十四畫 —

臺北 (台北) ........ 1

with upLaTeX & pxbabel



## Index

— Symbols —

€ ........ 1

3.14159265 ........ 1

— I —

İstanbul ........ 1

— S —

São Paulo ........ 1

— Α —

Αθήνα ........ 2

— Θ —

Θεσσαλονίκη ........ 2

— К —

Київ ........ 1

— М —

Москва ........ 1

— さ —

さいたま ........ 1

— と —

東京 ........ 1

— 사 —

서울 ........ 1

— 파 —

평양(平壤) ........ 1

— 匕部 —

北京 ........ 1

— 至部 —

臺北(台北) ........ 1

— द —

दिल्ली ........ 2

— म —

मुंबई ........ 2

— ก —

กรุงเทพมหานคร ........ 2

— น —

นนทบุรี ........ 2

— ا —

2 ........ أبو ظبي

— د —

2 ........ دبي

— י —

2 ........ ירושלים

— ת —

2 ........ תל אביב

with XeLaTeX & polyglossia

# Benchmark

| | makeindex | mendex | upmendex | xindy |
|---|---|---|---|---|
| internal encoding | 8bit 1byte | EUC-JP | UTF-16 | Unicode |
| Collator | locale | ASCII, Kana | ICU collator | |
| Latin | ✓ | ✓ ASCII | Lang:37, Locale:63 | Lang:33 |
| Greek | | | Lang:1, Locale:1 | Lang:1 |
| Cyrillic | | | Lang:9, Locale:9 | Lang:6 |
| Chinese | | | Lang:1, Locale:4 | |
| Japanese | | ✓ (Yomi, Dict) | ✓ (Yomi, Dict)<br>Lang:1, Locale:2 | |
| Korean | | | Lang:1, Locale:3 | |
| South Asian | | | Script:10, Lang:13, Locale:16 | |
| Thai / Lao | | | Script:2, Lang:2, Locale:2 | |
| Arabic | | | Lang:7, Locale:8 | |
| Hebrew | | | Lang:2, Locale:2 | Lang:1 |
| Other | | | | Lang:4 |
| Total | | Lang:2 | Script:22, Lang:72, Locale:109 | Lang:44 |

# Languages by Number of Native Speakers

| | Language | Script | speakers | ICU | polyglossia | upmendex | xindy |
|---|---|---|---|---|---|---|---|
| 1 | Chinese | Hanzi | 1,370,000,000 | ✓ | ✓ | ✓ | |
| 2 | English | Latin | 530,000,000 | ✓ | ✓ | ✓ | ✓ |
| 3 | Hindi | Devanagari | 490,000,000 | ✓ | ✓ | ✓ | |
| 4 | Spanish | Latin | 420,000,000 | ✓ | ✓ | ✓ | ✓ |
| 5 | Arabic | Arabic | 230,000,000 | ✓ | ✓ | ✓ | |
| 6 | Bengali | Bengali | 220,000,000 | ✓ | ✓ | ✓ v1.20 | |
| 7 | Portuguese | Latin | 215,000,000 | ✓ | ✓ | ✓ | ✓ |
| 8 | Russian | Cyrillic | 180,000,000 | ✓ | ✓ | ✓ | ✓ |
| 9 | Japanese | Kana & Hanzi | 134,000,000 | ✓ | ✓ | ✓ | |
| 10 | German | Latin | 130,000,000 | ✓ | ✓ | ✓ | ✓ |
| 11 | French | Latin | 123,000,000 | ✓ | ✓ | ✓ | ✓ |
| 12 | Punjabi | Gurmukhi | 90,000,000 | ✓ | ✓ | ✓ v1.20 | |
| 13 | Javanese | Latin | 75,000,000 | ✓ | ✓ | ✓ | |
| 14 | Korean | Hangul | 75,000,000 | ✓ | ✓ | ✓ | |
| 15 | Vietnamese | Latin | 70,000,000 | ✓ | ✓ | ✓ | ✓ |
| 16 | Telugu | Telugu | 70,000,000 | ✓ | ✓ | ✓ v1.20 | |
| 17 | Marathi | Devanagari | 68,000,000 | ✓ | ✓ | ✓ | |
| 18 | Tamil | Tamil | 74,000,000 | ✓ | ✓ | ✓ v1.20 | |
| 19 | Persian | Arabic | 46,000,000 | ✓ | ✓ | ✓ | |
| 20 | Urdu | Arabic | 61,000,000 | ✓ | ✓ | ✓ | |

Ref. https://ja.wikipedia.org/wiki/ネイティブスピーカーの数が多い言語の一覧

# Languages used on the Internet

| | Language | share | Script | ICU | pg | upm | xnd |
|---|---|---|---|---|---|---|---|
| 1 | English | 63.4 % | Latin | root | ✓ | ✓ | ✓ |
| 2 | Russian | 7.1 % | Cyrillic | ru | ✓ | ✓ | ✓ |
| 3 | Spanish | 3.9 % | Latin | es | ✓ | ✓ | ✓ |
| 4 | German | 3.7 % | Latin | de | ✓ | ✓ | ✓ |
| 5 | Turkish | 3.5 % | Latin | tr | ✓ | ✓ | ✓ |
| 6 | Persian | 2.5 % | Arabic | fa | ✓ | ✓ | |
| 7 | French | 2.0 % | Latin | root | ✓ | ✓ | ✓ |
| 8 | Japanese | 1.9 % | Kana | ja | ✓ | ✓ | |
| 9 | Portuguese | 1.8 % | Latin | pt | ✓ | ✓ | ✓ |
| 10 | Chinese | 1.3 % | Hanzi | zh | ✓ | ✓ | |
| 11 | Vietnamese | 1.3 % | Latin | vi | ✓ | ✓ | ✓ |
| 12 | Italian | 1.0 % | Latin | root | ✓ | ✓ | ✓ |
| 13 | Arabic | 0.9 % | Arabic | ar | ✓ | ✓ | |
| 14 | Polish | 0.9 % | Latin | pl | ✓ | ✓ | ✓ |
| 15 | Greek | 0.7 % | Greek | el | ✓ | ✓ | ✓ |
| 16 | Dutch | 0.7 % | Latin | nl | ✓ | ✓ | ✓ |
| 17 | Indonesian | 0.7 % | Latin | root | ✓ | ✓ | |
| 18 | Korean | 0.6 % | Hangul | ko | ✓ | ✓ | |
| 19 | Czech | 0.4 % | Latin | cs | ✓ | ✓ | ✓ |
| 20 | Thai | 0.4 % | Thai | th | ✓ | ✓ | |
| 21 | Ukrainian | 0.3 % | Cyrillic | uk | ✓ | ✓ | ✓ |
| 22 | Hebrew | 0.3 % | Hebrew | he | ✓ | ✓ | ✓ |
| 23 | Swedish | 0.3 % | Latin | sv | ✓ | ✓ | ✓ |
| 24 | Romanian | 0.3 % | Latin | ro | ✓ | ✓ | ✓ |
| 25 | Hungarian | 0.3 % | Latin | hu | ✓ | ✓ | ✓ |
| 26 | Danish | 0.2 % | Latin | da | ✓ | ✓ | ✓ |
| 27 | Slovak | 0.2 % | Latin | sk | ✓ | ✓ | ✓ |
| 28 | Serbian | 0.2 % | Latn, Cyrl | sr | ✓ | ✓ | ✓ |
| 29 | Bulgarian | 0.1 % | Cyrillic | bg | ✓ | ✓ | ✓ |
| 30 | Finnish | 0.1 % | Latin | fi | ✓ | ✓ | ✓ |
| 31 | Croatian | 0.1 % | Latin | hr | ✓ | ✓ | ✓ |
| 32 | Lithuanian | 0.1 % | Latin | lt | ✓ | ✓ | ✓ |
| 33 | Norwegian (Bokmål) | 0.1 % | Latin | nb | ✓ | ✓ | ✓ |
| 34 | Hindi | 0.1 % | Devanagari | hi | ✓ | ✓ | |
| 35 | Norwegian (nynorsk) | 0.1 % | Latin | nn | ✓ | ✓ | ✓ |
| 36 | Slovenian | 0.1 % | Latin | sl | ✓ | ✓ | ✓ |
| 37 | Latvian | 0.1 % | Latin | lv | ✓ | ✓ | ✓ |
| 38 | Estonian | 0.1 % | Latin | et | ✓ | ✓ | ✓ |
| 39 | Azerbaijani | < 0.1 % | Latin | az | | ✓ | |
| 40 | Catalan | < 0.1 % | Latin | root | ✓ | ✓ | |

Ref. https://ja.wikipedia.org/wiki/インターネットにおける言語の使用

# To Do

- Support more scripts
  - Khmer, Myanmar (Burmese)
  - Tibetan, Mongolian
  - Armenian, Georgian
  - Ethiopic (Amharic)
  - etc.
- Support more locales
  - Latin Script: Sorbian, Hausa, Igbo, Yoruba, Kalaallisut, Breton, Uzbek
  - etc.

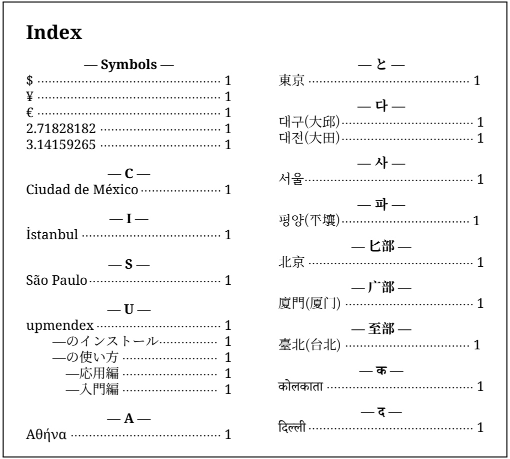
## Feedback is welcome

https://github.com/t-tk/upmendex-package/issues

# Summary

I introduced multilingual index processor **upmendex**.

- Feature
- Localization: 72 Languages, 22 Scripts
  - Latin, Cyrillic, Greek
  - CJK (Chinese, Japanese, Korean)
  - Devanagari, Bengali, Thai, ⋯
  - Arabic, Hebrew
  - Symbol, Number
- Multilingualization
  - Environment for upLaTeX/babel & XeLaTeX/polyglossia



## Index

— **Symbols** —

$ 1
¥ 1
€ 1
2.71828182 1
3.14159265 1

— **C** —

Ciudad de México 1

— **I** —

İstanbul 1

— **S** —

São Paulo 1

— **U** —

upmendex 1
—のインストール 1
—の使い方 1
—応用編 1
—入門編 1

— **A** —

Αθήνα 1

— **と** —

東京 1

— **다** —

대구(大邱) 1
대전(大田) 1

— **사** —

서울 1

— **파** —

평양(平壤) 1

— **匕部** —

北京 1

— **广部** —

廈門(厦门) 1

— **至部** —

臺北(台北) 1

— **क** —

कोलकाता 1

— **द** —

दिल्ली 1

# References

1. ASCII Nihongo TeX (Publishing TeX), ASCII MEDIA WORKS (web site by DWANGO Co., Ltd.). The site distributes mendexk source files.
2. Source/Document distribution of upmendex — multilingual index processor @ GitHub.
   upmendex @ CTAN
3. upTeX, upLaTeX — unicode version of pTeX, pLaTeX
4. International Components for Unicode (ICU)
5. PXbase — LaTeX: Support library for other PX packages @ GitHub. The repository distributes pxbabel.
   pxbase @ CTAN
6. polyglossia — An alternative to Babel for XeLaTeX and LuaLaTeX @ GitHub.
   polyglossia @ CTAN
7. “Indexing Makes Your Book Perfect” by SHIKANO Keiichiro at TUG2013, October, Tokyo.