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

田中 琢爾 TANAKA Takuji

2023年3月4日

田中 琢爾 TANAKA Takuji

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

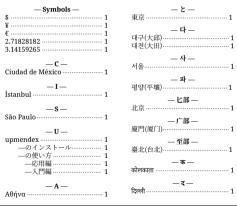
2023年3月4日 1/36

Overview

• Feature of **upmendex**

- Multilingual index processor
- Localization
 - Latin, Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, Thai
 - Arabic, Hebrew
- Multilingual environment
- Benchmark

Index



(a)

Feature of upmendex (Ver. 1.07)

Index processor

- Upper compatible with MakeIndex/Mendex
- Work with upLaTeX/LuaLaTeX/XeLaTeX
- Localization
 - Support 60 Languages / 12 Scripts
 - Latin (incl. non-English), Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, 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
	Hanzi / Pinyin	zh
Chinese	Hanzi / Stroke	zh-u-co-stroke
Chinese	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
Thai	Thai	th
Persian	Arabic	fa
Arabic	Arabic	ar
Hebrew	Hebrew	he
Yiddish	Hebrew	yi
Common	Symbol	
common	Number	

upmendex supports 60 languages, 12 scripts & 95 locales.

(日)

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

Index

ad

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				
AE1				
af 1				
AF 1				

default

phonebook sort order

(a)

AF------ 1

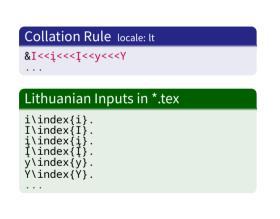
_ A _

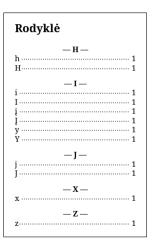
AD

中	琢爾	TANAKA	Takuji
---	----	--------	--------

2023年3月4日 6/36

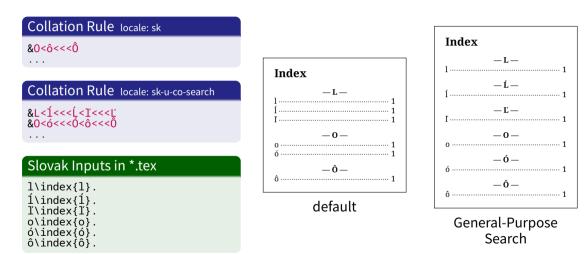
Lithuanian, Sort Order of Y





イロト イヨト イヨト イヨト

Slovak, Diacritical Mark

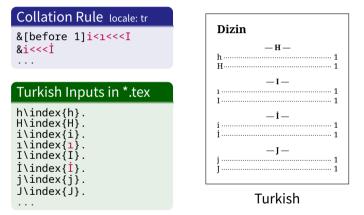


2023年3月4日 8/36

イロト イポト イヨト イヨト

Turkish, Dotless / Dotted I

language	upper	lower
Turkish	İ I	i
English	I	i



イロト イヨト イヨト

Hungarian, Digraphs and Trigraph

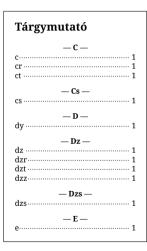
Collation Rule locale: hu

&C<cs<<Cs &D<dz<<Cz<CZ &DZ<dz<<DZ &DZ<dzs<<CZZ

. . .

Hungarian Inputs in *.tex

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



イロト イポト イヨト イヨト

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}.



Russian, Cyrillic



Greek

イロト イポト イヨト イヨト

2023年3月4日 11/36

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, 厂XY)} 鳥\index{鳥 (11, 鳥, niǎo, ろー幺^{*})} 風\index{風 (9, 風, fēng, 匸∠)} 月\index{月 (4, 月, yuè, 凵せ^{*})}

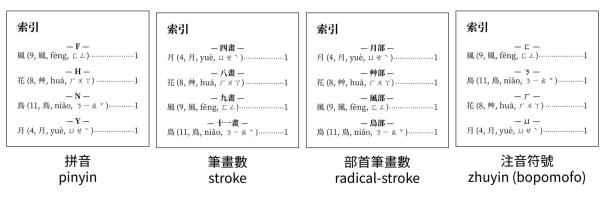
sort orde	locale	
Pinyin	拼音	zh
Stroke	筆畫數	zh-u-co-stroke
Radical-Stroke	部首筆畫數	zh-u-co-unihan
Zhuyin (Bopomofo)	注音符號	zh-u-co-zhuyin

Chinese, Han Ideograph Sort Order

Pinyin Sort Order 拼音 locale: zh	Radical-Stroke Sort Order 部首筆畫數 zh-u-co-unihan		
\centerline{\bfseries F}\par\nobreak \item 風 (9, 風, <mark>fēng</mark> , 匸乙)\leaders\hfill 1	\centerline{\bfseries 月部}\par\nobreak \item 月 (4, 月, yuè, 山せ`)\leaders\hfill 1		
\indexspace	\indexspace		
\centerline{\bfseries H}\par\nobreak \item 花 (8, 艸, huā, 厂XY)\leaders\hfill 1	\centerline{\bfseries 艸部}\par\nobreak \item 花 (8, 艸, huā, 厂X丫)\leaders\hfill 1		
Stroke Sort Order 筆畫數 zh-u-co-stroke	Zhuyin (Bopomofo) Sort Order 注音符號 zh-u-co-zhuyin		
Stroke Sort Order 筆畫數 zh-u-co-stroke \centerline{\bfseries 四畫}\par\nobreak \item 月 (4, 月, yuè, 山せ`)\leaders\hfill 1	Zhuyin (Bopomofo) Sort Order 注音符號 zh-u-co-zhuyin \centerline{\bfseries こ}\par\nobreak \item 風 (9, 風, fēng, こと)\leaders\hfill 1		
\centerLine{\bfseries 四畫}\par\nobreak	\bfseries		

upmendex Output *.ind

Chinese, Han Ideograph Sort Order



2023年3月4日 15/36

イロト イポト イヨト イヨト 二日

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"



拼音 pinyin



注音符號 zhuyin (bopomofo)

イロト イポト イヨト イヨト 二日

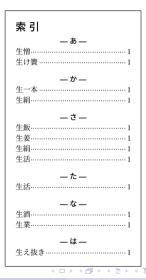
2023年3月4日 16/36

Japanese, Sort by Reading (Yomi)

Japanese Inputs with Reading (Yomi) in *.tex
$\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $
<pre>%% \index{reading@index_word} 生酒\index{なまざけ@生酒}。 生一本\index{さいっぽん@生一本}。 生け簀\index{いけす@生け簀}。 生絹\index{さぎぬ@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}}。</pre>
•••

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

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



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

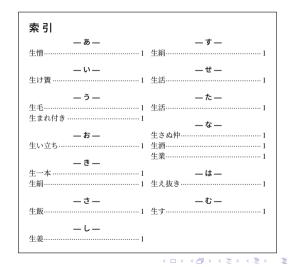
 生一本
 きいっぽん

 生け簀
 いけす

 生絹
 きぎぬ

 生え抜き
 はえぬき

Implemented by ASCII mendex.



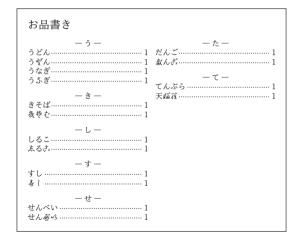
Japanese, Hentaigana

Inputs with Hentaigana in *.tex

後校で\index{後校で} うふぎ\index{うふぎ} えるよ\index{太るよ} 考し\index{表し} 天處雇\index{天處雇}

Dictionary for Hentaigana *.dic

index_word	reading
丧	き
授	そは
~ か	な
ふえむ	Ũ
さ	٢



イロト イポト イヨト イヨト

Extended Kana in JIS X 0213

Extended Kana Inputs in *.tex

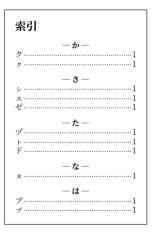
が\index{が}. % NGA ぎ\index{ぎ}. % NGI ぐ\index{ぐ}. % NGU ... index{ゟ}. % Hiragana Digraph Yori つ\index{つ}. % Katakana Digraph Koto

Aynu itak Inputs in *.tex

ク\index{ク} シ\index{シ} ス\index{ス プ\index{プ] ピ\index{ピ] ツ\index{ピ ド\index{ド}

. . .

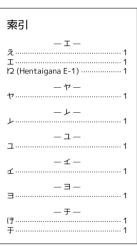


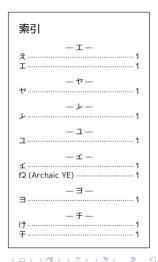


イロト イポト イヨト イヨト 二日

Japanese, Archaic Kana







田中 琢爾 TANAKA Takuji

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

Korean, Modarn / Archaic Hangul



	Unicode block	style		upmendex	upLaTeX	XeLaTeX
modarn	Hangul Syllables	composed	완성형	\checkmark	\checkmark	\checkmark
ex. 일	Hangul Jamo	decompopsed	조합형	\checkmark	N.A.	\checkmark
archaic	Private Use Area	composed	완성형	via dictionary	\checkmark	\checkmark
ex. ·읦	Hangul Jamo	decomposed	조합형	\checkmark	N.A.	\checkmark

・ロト ・ 日下・ ・ ヨト・

Korean Hangul

Hangul Inputs in *.tex

쓰\index{쓰 (composed)}. ~−\index{~− (decomposed)}.

```
☆ ·\index{☆ · (archaic)}.
☆ ·:\index{☆ ·: (archaic with tone mark)}.
☆\index{씃 (Hanyang PUA)}.
```

Dictionary for PUA code *.dic

Hanyang PUA code	decomposed
м.	м.
씻	ル .1

찾아보기
_ ^ _ 스 (composed)1
쓰 (composed)······1 쓰 (decomposed) ······1 ·쏘 (archaic with tone mark) ·····1
や (archaic)
_ ~ _
_ * _ _ (archaic)1
_ * _ _ (archaic)1
_ ^

Complex Text Layout

- Devanagari, Thai
- Arabic, Hebrew: R-to-L typeset
- Symbol, Number

Devanagari & Thai (experimental)

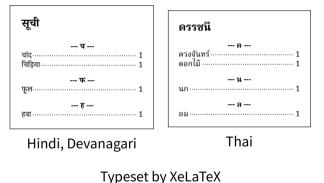
Hindi Inputs in *.tex

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 }



•••

田中 琢爾 TANAKA Takuji

2023年3月4日 25/36

3

Arabic & Hebrew (experimental)

Arabic Inputs in *.tex

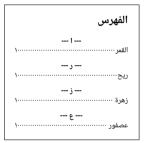
زهرة } \index زهرة } { عصفور } \index عصفور { ريح } \index ريح القمر } القمر }

...

Hebrew Inputs in *.tex

פָרַח {index פּרָח פּרַח ציפור {index ציפור { רוים {index רוים } רירח {index ירח }







Hebrew

医子宫 医子宫 医子宫的

.

ירח ירח

--- 5 ----

פרחפרח

--- ¥ ----

1 זיתור

---- 7 ----

1

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

מפתח

Symbol, Number

Script	charType	example	treatment by upmendex
Latin	Lu, Ll, Lo, : letters	ABCabc A a 🕘	directly pass to ICU collator
Greek	Lu, Ll, Lo, : letters	ΑΒΓαβγ	direct
Cyrillic	Lu, Ll, Lo, : letters	АБВабв	direct
Kana	Lo: other letter	あいうアア定ァ゚	direct
Hangul	Lo : other letter	가나다ᄁᆫᅞᆣᄳᆋ	direct
Hanzi	Lo : other letter	花鳥風月	lookup dictionary or direct
—	Lm: modifier letter	* o	direct
Number	Nd : dicimal digit number	012012	direct
	No : other number	¹² 3 4 567,(8)9.	lookup dictionary or direct
	Sk : modifier symbol	/ ^ ` v o	direct
Symbol	Sm: math symbol	÷⊳#	lookup dictionary or direct
	So : other symbol	⋬ѿ☎♥€∭	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	diect
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

(a)

Multilingual Environment with upLaTeX/pxbabel

Block Setting for Scripts in Style File *.ist

script_preamble script_postamble		"\n\\fontencoding{T2A}\\selectfont" "\n\\fontencoding{T1}\\selectfont"
script_preamble script_postamble	hangul hangul	"\n\\begin{otherlanguage}{korean}" "\n\\end{otherlanguage}"
script_preamble script_postamble	hanzi hanzi	<pre>"\n\\begin{otherlanguage}{tchinese}" "\n\\end{otherlanguage}"</pre>

upmendex Output *.ind

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

(I) < ((()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) < (()) <

Multilingual Environment with XeLaTeX/polyglossia

Block Setting for Scripts in Style File *.ist

hebrew

script_preamble
script_postamble

script_preamble script_postamble

script_preamble
script_postamble

script_preamble
script_postamble

```
cyrillic "\n\\end{russian}"
kana "\n\\begin{japanese}"
kana "\n\\end{japanese}"
hangul "\n\n\begin{korean}"
hangul "\n\\end{korean}"
hebrew "\n\begin{hebrew}"
```

cyrillic "\n\\begin{russian}"

"\n\\begin{hebrew}"
"\n\\end{hebrew}"

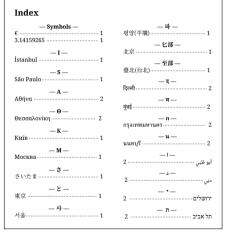
upmendex Output *.ind

```
\centerline{--- さ ---}\par\nobreak
 \item さいたま\leaders\hbox{~}\hfill {1}
\end{ japanese}
\begin{korean}
 \indexspace
\item 대구(大邱)\leaders\hbox{~}\hfill {1}
\end{korean}
\begin{hebrew}
 \indexspace
\centerline{--- א ---}\par\nobreak
\item \leaders\hbox{~}\hfill {2}
\end{hebrew}
```

Output of Multilingual Index



with upLaTeX & pxbabel



with XeLaTeX & polyglossia

Benchmark

	makeindex	mendex	upmendex	xindy		
internal encoding	8bit 1byte	EUC-JP	UTF-16	Unicode		
Collator	locale	ASCII, Kana	ICU collator			
Latin	\checkmark	√ ASCII	Lang:37, Locale:62	Lang:32		
Greek			Lang:1, Locale:1	Lang:1		
Cyrillic			Lang:9, Locale:9	Lang:6		
Chinese			🗍 Lāng:1, Locale:4 🗍			
Japanese		√ (Yomi, Dict)	√ (Yomi, Dict)			
		V (TOTTI, DICt)	Lang:1, Locale:2			
Korean			Lang:1, Locale:3			
Devanagari			Lang:3, Locale:3			
Thai			Lang:1, Locale:1			
Arabic			Lang:6, Locale:7			
Hebrew			Lang:2, Locale:3	Lang:1		
Other				Lang:4		
Total		Lang:2	Lang:60, Locale:95	Lang:44		

イロト イロト イヨト イヨ

Languages by Number of Native Speakers

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

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

イロト イヨト イヨト イヨ

< **D** > < **A**

Languages used on the Internet

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

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

▶ < 프 > < 프</p>

To Do

• Support more scripts

- Bengali, Telugu, Tamil, Malayalam, Kannada, Gujarati, Oriya, Sinhala
- Khmer, Lao, Myanmar (Burmese)
- Tibetan, Mongolian
- Armenian, Georgian
- Ethiopic (Amharic)
- etc.
- Support more locales
 - Latin Script: Sorbian, Hausa, Igbo, Yoruba, Kalaallisut, Breton, Uzbek
 - etc.
- Add style options
 - script_head

Feedback is welcome

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

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

Summary

I introduced multilingual index processor **upmendex**.

- Feature
- Localization: 60 Languages, 12 Scripts
 - Latin, Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, Thai, Arabic, Hebrew
 - Symbol, Number
- Multilingualization
 - Environment for upLaTeX/babel & XeLaTeX/polyglossia

Index



イロト イポト イヨト イヨト

References

- ASCII Nihongo TeX (Publishing TeX), ASCII MEDIA WORKS (web site by DWANGO Co., Ltd.). The site distributes mendexk source files.
- Source/Document distribution of upmendex multilingual index processor @ GitHub. upmendex @ CTAN
- upTeX, upLaTeX unicode version of pTeX, pLaTeX
- International Components for Unicode (ICU)
- PXbase LaTeX: Support library for other PX packages @ GitHub. The repository distributes pxbabel. pxbase @ CTAN
- polyglossia An alternative to Babel for XeLaTeX and LuaLaTeX @ GitHub. polyglossia @ CTAN
- ⁽⁰⁾ "Indexing Makes Your Book Perfect" by SHIKANO Keiichiro at TUG2013, October, Tokyo.