upmendex で作る多言語索引

Multilingual index processing by upmendex

田中琢爾
TANAKA Takuji

2022 年 11 月 19 日
TeXConf 2022

概要

upmendex は、索引作成ソフトウェア makeindex の日本語対応版 mendex をさらに Unicode 化した多言語拡張である。2014 年頃から開発をすすめ昨年正式版 ver 1.00 をリリースした。日中韓 (CJK) と欧文 (非英語を含むラテン文字、キリル文字、ギリシャ文字) に加え実験的ながらタイ文字、デヴァーナガリー、アラビア文字、ヘブライ文字をサポートした。ソートに ICU (International Components for Unicode) を用い、柔軟なカスタマイズが可能である。本講演では、upmendex を upLaTeX+babel や XeLaTeX+polyglossia と組み合わせた実例を示しながら多言語索引処理の詳細を解説する。

各スクリプトの扱い

ラテン文字、キリル文字、ギリシャ文字

ラテン文字、キリル文字、ギリシャ文字の言語では、ソート順の問題、ダイアクリティカルマーク・ダイグラフ/トライグラフの処理について述べる。

ラテン文字の言語では、英語のアルファベット順にそのまま従うとは限らず、言語や用途によりソート順が異なることがある。例えば、ドイツ語では電話帳順と呼ばれるソート順がある。リトアニア語では “Y” が “X” と “Z” の間ではなく “I” と同じ位置にソートされる。またダイアクリティカルマーク付きの文字を、独立した文字と見なすのか、ダイアクリティカルマーク無しの文字と同等の位置にソートするのか、多様である。さらに、言語によっては特定の二文字、三文字の組を一つの文字と見なすダイグラフ・トライグラフがある (例えばハンガリー語の “CS”, “DZ”, “DZS”)。

ICU ではこれらを locale の設定によりカスタマイズ出来る。upmendex では、ICU に locale の設定を渡してソートしさらにそのソート順に応じて適切な見出し項目を立てるように動作する。

CJK (日中韓)

中国語では、漢字のソート順をどうするかが一つの焦点になる。ICU ではピンイン順・画数順・部首画数順・注音符号順の 4 種類をサポートし locale の

設定で切り替え可能となっており、upmendex ではそれに応じて適切な見出しを付けるよう対応している。

日本語では、読み仮名を辞書順にソートする必要がある。upmendex では上流のソフトウェアである ASCII 社の mendex で実装された機能をそのまま継承し、読みを (u)pLaTeX 入力の段階で明示して入力する方法と辞書ファイルを用意し辞書から読みを得る方法をサポートしている。変体仮名は 2017 年に Unicode 10.0 で定義されたものの ICU のソートが未対応の状況だが、(up)mendex 側の辞書機能を用いることで適切に処理できる。upmendex では mendex の日本語機能に加え、JIS X 0213 で定義された鼻濁音「がぎぐげご」やアイヌ語の仮名「ㇰシストヌ…ㇷゼヅド」や最近 Unicode で定義された仮名「や行え段 “Ye”」等にも対応している。

韓国語では、ハングルの文字コード上の表現形式に次の二種類がある。すなわち、ハングル一文字を合成済みの文字コードで表現する方法 (完成型，composed) と音素の文字コードの組み合わせで表現する方法 (組合型，decomposed) である。現代語は全て完成型で表現されソフトウェア上の扱いは容易だが、古ハングルは Unicode 標準では組合型で表現する必要がある。upmendex では現代語に加え古ハングルの組合型にも対応している。

タイ文字、デーヴァナーガリー、アラビア文字、ヘブライ文字

upmendex では、タイ文字、デーヴァナーガリー、アラビア文字、ヘブライ文字も扱えるよう実装しているが、それらの言語に対する知見が開発者に乏しいため今のところ実験的と位置づけている。既に実用的な水準まで届いているつもりだが確信はなく、仕様の不備や不適切な点があればご教示願いたい。

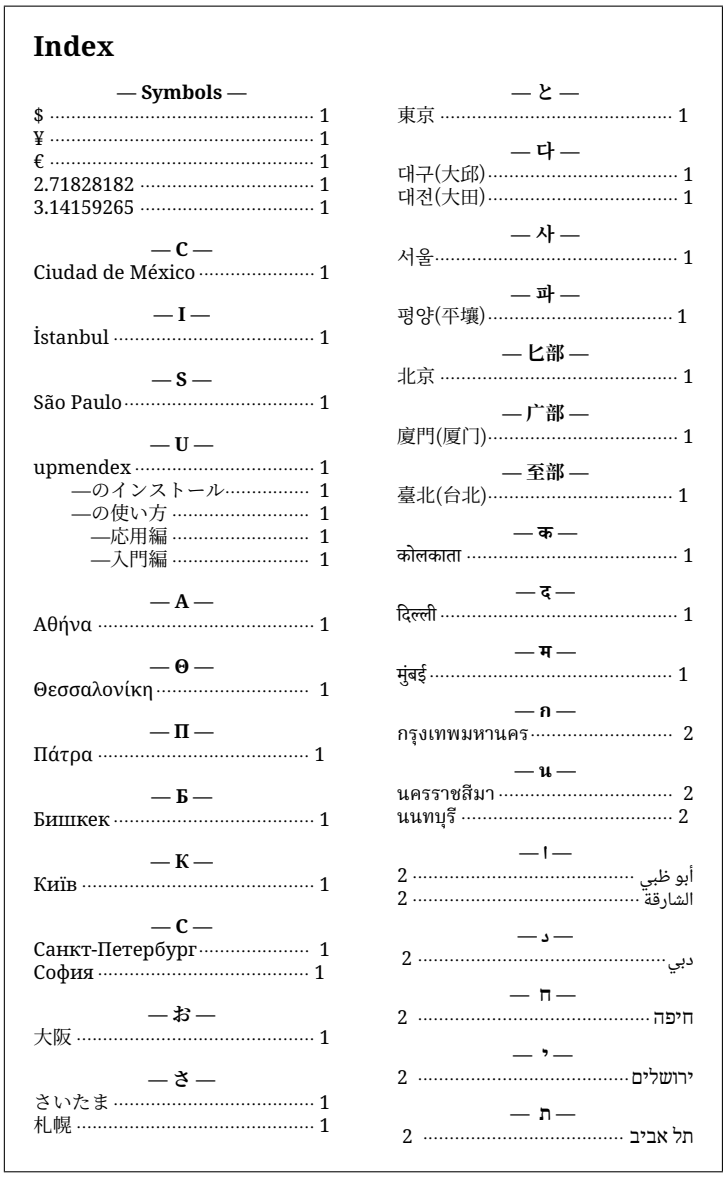
これらの文字はいずれも組版の際に複雑なレンダリング処理を要し、さらにアラビア文字、ヘブライ文字では右から左へ (R-to-L) 組む必要がある。upmendex では直接それらの処理には関与せず、文字コードの操作と後述のブロック毎の環境設定の機能のみ実装している。各スクリプト毎の環境は babel や polyglossia で用意することを想定している。

記号、数字

Unicode では、記号類や数字の種類も非常に多彩であるが、その用途の分類が “charType” で示されている。upmendex では目的に応じてソート順を制御できるようにするため、charType を参照して各種記号 (`So`, 「☃☜☎♥⚽㊊」等)、通貨記号 (`Sc`, 「€$¢£¥₩」等)、数学記号 (`Sm`, 「÷▷＃」等)、その他数字 (`No`, 「¹²③❹」等) については最初に辞書で読みの検索を試み、見つからない場合にはそのまま ICU collator に渡す。その他の記号類 (例えば句読点類 `Po`, 「?!¡¿†＃§¶」等) は、辞書を見ずに直接 ICU collator に渡す。

多言語索引

多言語環境として upLaTeX+babel や XeLaTeX+polyglossia と upmendex を組み合わせる場合、索引の中でも各スクリプト毎にフォントの設定などを切り替える必要がある。upmendex では `script_preamble`, `script_postamble` を style ファイル (.ist) 内で定義することにより各スクリプトのブロック毎に環境設定が出来るようにした。ここでは XeLaTeX+polyglossia と upmendex を組み合わせて組版した例を示す。



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

— Θ —
Θεσσαλονίκη 1

— Π —
Πάτρα 1

— Б —
Бишкек 1

— К —
Київ 1

— С —
Санкт-Петербург 1
София 1

— お —
大阪 1

— さ —
さいたま 1
札幌 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

— म —
मुंबई 1

— ก —
กรุงเทพมหานคร 2

— น —
นครราชสีมา 2
นนทบุรี 2

— ا —
أبو ظبي 2
الشارقة 2

— د —
دبي 2

— ח —
חיפה 2

— י —
ירושלים 2

— ת —
תל אביב 2

XeLaTeX+polyglossia+upmendex で作成した多言語索引の例

upmendex の現在と未来

upmendex は多言語索引作成のツールとして世界中の主要な言語に広く対応してきた。対応しているスクリプト (12 種) や言語 (60 言語, 95 locale) の種類の多さは同種のソフトウェアの中でもトップクラスであり、特に CJK のサポート内容は最も充実していると自負している。

しかし世界は広く、upmendex で未対応のスクリプトも、ICU では既対応だが upmendex では未対応の locale もまだ多く残っている。今後、スクリプトの追加 (ベンガル文字・テルグ文字・タミル文字等) に加え、locale をさらに拡充し世界制覇を目指したい。