pxcjkcat パッケージ

八登崇之 （Takayuki YATO; aka. “ZR”）

v1.4 [2022/06/06]

概要

upTEX エンジンには「ソース中の非 ASCII 文字の扱い」（和文・欧文の何れとして扱うか、等）を制御するための「和文カテゴリ（kcatcode）」という設定がある。本パッケージは、upTEX の和文カテゴリを扱うための LATEX の文書作成者向けのインタフェースを提供する。

目次

1 前提知識

1.1 和文カテゴリ

upTEX エンジンの「**和文カテゴリ**（kcatcode）」の設定は「ソース中の非 ASCII 文字の振舞」を制御するためのものである。例えば「ア」（U+30A2）という文字は普通は（既定値では）仮名として扱われるが、U+30A2 の和文カテゴリを変更することで、これを「和文記号扱い」「欧文扱い」などに変えることができる。

upTEX の仕様では和文カテゴリの値は 15〜19 の範囲の整数（**和文カテゴリコード**）として表される。本パッケージでは操作を直感的にするため和文カテゴリの値に名前（**カテゴリ ID**）を付けて扱う。和文カテゴリの値の一覧は以下の通りである。

コード	カテゴリ ID	意味
15	`noncjk`	欧文扱い
16	`kanji`	漢字扱い
17	`kana`	仮名扱い
18	`cjk`	「その他の和文」扱い
19	`hangul`	ハングル扱い

和文カテゴリの設定による実際の「振舞」の違いには以下のようなものがある[*1]。

- 和文カテゴリが `noncjk`（15）である文字は欧文として扱われる。対して、`noncjk` 以外の文字は和文扱いとなる。
  ※なお、欧文扱いとなる文字は欧文 LaTeX（pdfLaTeX）と同じ機構で処理される。すなわち、TEX のレベルでは UTF-8 のバイト列として扱われ、これを inputenc パッケージが適切に処理することで初めて「文字」として認識される。
- (u)pLaTeX には「行末にある文字が和文文字の場合には当該の改行は空白にならずに無視される」という独自の入力規則があるが、和文カテゴリが `hangul`（19）である和文文字が行末にある場合は欧文と同様に改行は空白を発生させる。
- LaTeX の \ から始まる命令名（制御綴）の構成について「複数文字の名前を作れるのは英字のみ」という規則がある。命令名構成の規則について、和文カテゴリが `kanji`（16）・`kana`（17）・`hangul`（19）の和文文字は「英字」と同じ扱いになり、`cjk`（18）の和文文字は「英字以外の欧文文字」と同じ扱いになる。

ただし、upTEX の仕様では、和文文字カテゴリの設定は“文字単位”ではなく“Unicode ブロック単位”で行うようになっている（一部例外あり）。従って、例えば「ア」（U+30A2）の和文カテゴリを変更したいという場合には、U+30A2 が属する“Katakana”（U+30A0〜30FF）のブロックについて和文カテゴリの設定を行うことになる。

[*1] なお、現在の upTEX の仕様では `kanji`（16）と `kana`（17）の間には「振舞」の違いは存在しないようである。

本パッケージが提供する命令を利用すると「“Katakana”の和文カテゴリを cjk（18）に変更する」という設定は

```
\cjkcategory{kana}{cjk}
```

で実現できる。

1.2 和文カテゴリモード

和文カテゴリモードは全てのブロックに対する和文カテゴリの一括設定（プリセット設定）のことである（TODO：ちゃんと説明する）。

モード設定には以下のものがある。何れのモードも、「CJK 中核セット」（後述）のブロック群の設定は共通で、その他のブロックが欧文扱い（noncjk）であるか和文扱い（cjk[*2]）であるかが異なる。

- forcecjk： 全てのブロックを和文扱い（noncjk 以外）とする。和文フォントの中の Unicode 値の割当がある全ての文字を和文文字として直接用いることができる。
- default： 現在のモード CCV に対応する upTEX の版（例えばモード CCV が 3 ならば v1.23）における既定設定と一致させる。
  ※モード CCV が 2 以下の場合は forcecjk と同一の設定になる。
- prefercjk： 和文扱いのブロックとして「CJK 中核セット」の他に「Adobe の定める CJK 文字集合[*3]の何れかと共通部分をもつ文字ブロック」を加えて、残りを欧文扱いに設定する。
- prefercjkvar：prefercjk からギリシャ・キリル文字のブロックを欧文扱いに変更したもの。
- japanese： 和文扱いのブロックとして「CJK 中核セット」の他に「Adobe-Japan1 の全角幅のグリフの何れかと共通部分をもつ文字ブロック」を加えて、残りを欧文扱いに設定する。
- japanesevar： japanese からギリシャ・キリル文字のブロックを欧文扱いに変更したもの。
- prefernoncjk：「CJK 中核セット」のブロックのみを和文扱いとし、残りを欧文扱いにする。

※各モードでの具体的な設定値については 6 節を参照。

■CJK 中核セット 「CJK 中核セット」は以下の文字種が属するブロックのセットを指す。（括弧内は、モード設定において当該のブロックに設定される和文カテゴリの値。）

- 漢字・部首・注音字母（kanji）
- ひらがな・カタカナ（kana）
- CJK 記号の一部・全角半角互換形・彝文字・西夏文字・女書文字・契丹文字（cjk）
  ※モード CCV が 2 以上の場合、cjk12 の再分割の cjk1b、cjk1c は kana に設定される。
- ハングル完成形・ハングル字母（hangul）

[*2] 全てのモード設定において、「CJK 中核セット」**以外**のブロックのカテゴリは必ず noncjk か cjk の何れかになる。
[*3] Adobe-Japan1、Adobe-GB1、Adobe-CNS1、Adobe-Korea1 の 4 つ。

1.3 和文トークン状態

upTEX には、和文カテゴリ（kcatcode）とは別に、非 ASCII 文字全体の和文・欧文扱いの設定を一斉に切り替えるための命令が存在する[*4]。

- `\enablecjktoken`： 和文・欧文扱いの設定を和文カテゴリの設定に従わせる。
- `\disablecjktoken`： 和文カテゴリ設定に関わらず非 ASCII 文字全体を欧文扱いにする。
  ※あたかも全ブロックの和文カテゴリを `noncjk`（15）に設定したのと同じ状態になる。
- `\forcecjktoken`： 和文カテゴリ設定に関わらず非 ASCII 文字全体を和文扱いにする。
  ※和文カテゴリが `noncjk`（15）であるブロックはあたかもそれが `cjk`（18）であるように動作する。

これらの命令群により変更されるパラメタのことを本パッケージでは「**和文トークン状態**」と呼ぶことにする[*5]。

2 パッケージ読込

```
\usepackage[<オプション>]{pxcjkcat}
```

オプションとして以下のものが指定できる。

- **モード CCV 指定**：「モード CCV」の値を指定するオプション。
  ※モード CCV については 4 節を参照。
  ※モード CCV の既定値は 1 であり、これは極めて古い版と互換にすることを意味する。モード設定（`prefernoncjk` 等）を利用する場合には、適切なモード CCV のオプションを指定するのが望ましい。
  - `ccv1`： モード CCV を 1（upTEX v0.11〜0.28 と互換）とする。**既定値。**
  - `ccv2`： モード CCV を 2（upTEX v0.29〜1.22 と互換）とする。
  - `ccv3`： モード CCV を 3（upTEX v1.23 と互換）とする。
  - `ccv4`： モード CCV を 4（upTEX v1.25 以降と互換）とする。
  - `real` または `ccv+`： モード CCV を upTEX の実際の CCV と一致させる。
- **和文カテゴリモード値**： `\cjkcategorymode` 命令で有効なモード値（`prefernoncjk` 等）をオプションとして指定可能で、この場合、和文カテゴリがモードに従って設定される。
  ※和文カテゴリモード値オプションが指定されていない場合は、パッケージ読込時に和文カテゴリが変更されることはない。
- `nomode`： 和文カテゴリモード値オプションの効果を打ち消す。

[*4] これらの命令は本パッケージが提供するものではなく upTEX に元から存在するものであることに注意。
[*5] 「和文トークン状態」を指す公式の用語は存在しないと思われる。

3 機能

3.1 和文カテゴリ設定に関連する機能

- `\cjkcategory{`⟨ブロック⟩`,...}{`⟨カテゴリ⟩`}`：⟨ブロック⟩で指定される文字ブロック（複数指定が可能）の和文カテゴリを⟨カテゴリ⟩に変更する。
  ⟨ブロック⟩は以下の何れかの形式で指定する：
  - ブロック ID（5 節参照）[*6]
  - 非 ASCII 文字 1 つ： 当該の文字が属するブロックを指す。
    ※「文字の属するブロック」は、モード CCV 設定とは無関係であり常に upTEX の実際のブロック定義に従う。従ってその動作は upTEX の版に依存することに注意。
  - 符号値（整数値）： 当該の符号値の文字が属するブロックを指す。符号値は以下の形式で指定できる：
    * ⟨16 進表記⟩： 例えば `1F600`。
    * `"`⟨16 進表記⟩： 例えば `"1F600`。
      ※ 16 進数字の `A`〜`F` は大文字で書く。
    * `+`⟨整数⟩： 例えば `+128512` は 10 進表記で 128512、すなわち U+1F600 を表す。
      ※ ⟨整数⟩ の部分には一般に任意の「(LA)TEX で整数を表すテキスト」が記述できる。例えば `+\value{mycode}` のように書くとカウンタ値を指定できる。
    ※前項と同様、動作が upTEX の版に依存することに注意を要する。

  ⟨カテゴリ⟩は「カテゴリ ID」または「カテゴリコード」（括弧内の整数値）で指定する。
  - `noncjk`（15）： 欧文扱い
  - `kanji` または `han`（16）： 漢字扱い
  - `kana`（17）： 仮名扱い
  - `cjk`（18）：「その他の和文」扱い
  - `hangul`（19）： ハングル扱い

  和文カテゴリの変更は局所的（グルーピングに従う）である。
  ※モード CCV の指定は `\cjkcategory` の動作には影響を与えない。
  ※“Basic Latin”ブロック（`latn`）のカテゴリは常に `noncjk` でなければならず、`noncjk` 以外に変更しようとするとエラーになる。
- `\cjkcategorymode{`⟨モード⟩`}`： 全てのブロックの和文カテゴリの一括設定（モード設定）を行う。有効なモード設定の値は以下の通りである。
  ※モード設定の詳細については 1.2 節を参照。

[*6] ブロック ID による指定は upTEX の版の影響を受けない。例えば、upTEX の版（およびモード CCV 指定）が何であっても、`latn1` は常に“Latin-1 Supplement”の範囲（U+0080〜00FF）を指す。実 CCV が 3 以上の場合に `latn1` のカテゴリ設定を行った場合は、実際には `latnx` と `latny` の 2 つのブロックに対して設定が行われる。

  - forcecjk
  - default
  - prefercjk
  - prefercjkvar
  - japanese
  - japanesevar
  - prefernoncjk

3.2 和文トークン状態に関する機能

※「和文トークン状態」については 1.3 節を参照。

- \getcjktokenmode： 現在の和文トークン状態の取得して、それを表す整数値を \thecjktokenmode に設定する。
  - 0： \disablecjktoken の状態。
  - 1： \enablecjktoken の状態。
  - 2： \forcecjktoken の状態。
  - 255： 状態取得に失敗した[*7]。
- \setcjktokenmode{⟨整数値⟩}： \getcjktokenmode の規則の整数値を用いて和文トークン状態を設定する。
- \withcjktokendisabled{⟨コード⟩}：一時的に \disablecjktoken に変更した状態で、⟨コード⟩ を実行する。
- \withcjktokenenabled{⟨コード⟩}： 一時的に \enablecjktoken に変更した状態で、⟨コード⟩ を実行する。
- \withcjktokenforced{⟨コード⟩}： 一時的に \forcecjktoken に変更した状態で、⟨コード⟩ を実行する。
  ※以上 3 つの命令はどれも、「和文トークン状態を設定して ⟨コード⟩ を実行した後和文トークン状態を再設定する」という動作を行う。そのため新たに局所化グループに入ることはない。

4 文字分類バージョン（CCV）

注意： 4 節および 5 に述べられている説明は現状の仕様と食い違っている部分があり、大幅な改訂が必要な状態である。少なくとも 1.1 版以降の仕様では、モード CCV の影響を受けるのはモード設定の定義のみであり、ブロック分割は常に実際のエンジンのものに一致させている。

[*7] TEX Live 2022 以降で ε-TEX 拡張無しの upTEX エンジンの場合は取得ができない。しかし、2014 年以降の LATEX カーネルは ε-TEX 拡張を必須としているので、そのようなエンジンの上で LATEX が動作している可能性はほぼ考えられない。従って、現実的には 255 が返ることはないと思ってよい。

upTEX エンジンでの文字ブロックの分割および各ブロックの和文カテゴリの既定値は改版時に変更され、これが互換性の問題を起こす可能性がある。本パッケージでは、パッケージの機能を用いて設定された和文カテゴリの値がエンジンの改版により変化することを防ぐため、「特定のエンジンのバージョンを指定して、その動作を模倣する」という方針をとる。

4.1 文字分類バージョンとは何か

- 文字ブロックの分割の違いを「**文字分類バージョン**（CCV；Character Category Version）」と呼ぶことにする。現状では次のものが存在する[*8]。
  - バージョン 1： upTEX v0.11〜0.28 と互換
  - バージョン 2： upTEX v0.29〜1.22 と互換
  - バージョン 3： upTEX v1.23 と互換
  - バージョン 4： upTEX v1.25 以降と互換
- pxcjkcat の読込時に、そのオプションにおいて「その文書が依拠する CCV の値」（これを**モード CCV** と呼ぶ）を指定する。すなわち、オプション `ccv`N（$N = 1$〜3）を指定すると、モード CCV が N になる。
- モード CCV を使用するエンジンの実の CCV と常に一致させたい場合は、`ccv+` というオプションを指定すればよい。ただしこの場合は当然、和文カテゴリ設定がエンジンの版に依存することになる。
- モード CCV の既定値は 1（`ccv1`）である。この場合、ブロック分割の状態は Unicode[*9]のブロック定義と完全に一致する。
- モード設定で `default` を指定した場合は、和文カテゴリの設定は「モード CCV に対応するエンジンの版の既定値」に一致する。他のモード設定の実際の設定値も `default` を基礎にして決まるので、モード CCV により多少の差異が出る。

4.2 注意事項

- Unicode の改版による「文字ブロックの追加」については「当該の版のエンジンで未対応の文字ブロックの文字は決して使われない」ことを仮定すれば互換性を損なうことがないため、特に対策を行わない。
  従って、エンジンの改版が「文字ブロックの追加」だけを伴う場合は、それは新しい CCV とは見なされない。例えば、v1.00 → v1.10 の改修では幾つかのブロックが追加されたが、CCV は 2 のままである。
- もちろん、「旧版の動作の模倣」は本パッケージの機能を用いた場合に限られ、upTEX エンジ

[*8] upTEX の v1.24 には文字ブロック分割に関してバグが存在するので、CCV の定義からは除外する。なお、実際に v1.24 の upTEX で本パッケージが読み込まれた場合は、実 CCV は 1 と見なされる（はずである）。
[*9] エンジンの版に対応する版の Unicode。

ンの和文カテゴリコード（kcatcode）の処理自体は何も変更されない。また、本パッケージの機能を用いる以外の方法（エンジン既定のままの場合を含む）で設定された和文カテゴリ値については、当然、エンジンの版による差異が生じうる。

5 文字ブロック

5.1 Unicode ブロックの一覧

以下は、Unicode が定めるブロックと直接に対応するブロックの一覧である。この表の「ブロック ID」の欄が、`\cjkcategory` 命令で指定するブロック ID を示す。「CCV」の欄は、そのブロックがサポートされる実 CCV の範囲の下限を表す。[*10]

ブロックID	CCV	符号値範囲	ブロック名称
latn	1	U+ 0000 〜 007F	Basic Latin
latn1	1	U+ 0080 〜 00FF	Latin-1 Supplement
latnA	1	U+ 0100 〜 017F	Latin Extended-A
latnB	1	U+ 0180 〜 024F	Latin Extended-B
latn2	1	U+ 0250 〜 02AF	IPA Extensions
sym01	1	U+ 02B0 〜 02FF	Spacing Modifier Letters
sym02	1	U+ 0300 〜 036F	Combining Diacritical Marks
grek	1	U+ 0370 〜 03FF	Greek and Coptic
cyrl	1	U+ 0400 〜 04FF	Cyrillic
cyrl1	1	U+ 0500 〜 052F	Cyrillic Supplement
armn	1	U+ 0530 〜 058F	Armenian
hebr	1	U+ 0590 〜 05FF	Hebrew
arab	1	U+ 0600 〜 06FF	Arabic
syrc	1	U+ 0700 〜 074F	Syriac
arab1	1	U+ 0750 〜 077F	Arabic Supplement
thaa	1	U+ 0780 〜 07BF	Thaana
nkoo	1	U+ 07C0 〜 07FF	NKo
samr	2	U+ 0800 〜 083F	Samaritan
mand	2	U+ 0840 〜 085F	Mandaic
syrc1	3	U+ 0860 〜 086F	Syriac Supplement
arabA	2	U+ 08A0 〜 08FF	Arabic Extended-A
deva	1	U+ 0900 〜 097F	Devanagari
beng	1	U+ 0980 〜 09FF	Bengali
guru	1	U+ 0A00 〜 0A7F	Gurmukhi
gujr	1	U+ 0A80 〜 0AFF	Gujarati
orya	1	U+ 0B00 〜 0B7F	Oriya
taml	1	U+ 0B80 〜 0BFF	Tamil
telu	1	U+ 0C00 〜 0C7F	Telugu
knda	1	U+ 0C80 〜 0CFF	Kannada
mlym	1	U+ 0D00 〜 0D7F	Malayalam
sinh	1	U+ 0D80 〜 0DFF	Sinhala
thai	1	U+ 0E00 〜 0E7F	Thai

[*10] これは参考情報であり、現状の仕様ではあまり意味をもたない。

laoo	1	U+	0E80 ~	0EFF	Lao
tibt	1	U+	0F00 ~	0FFF	Tibetan
mymr	1	U+	1000 ~	109F	Myanmar
geor	1	U+	10A0 ~	10FF	Georgian
hang1	1	U+	1100 ~	11FF	Hangul Jamo
ethi	1	U+	1200 ~	137F	Ethiopic
ethi1	1	U+	1380 ~	139F	Ethiopic Supplement
cher	1	U+	13A0 ~	13FF	Cherokee
cans	1	U+	1400 ~	167F	Unified Canadian Aboriginal Syllabics
ogam	1	U+	1680 ~	169F	Ogham
runr	1	U+	16A0 ~	16FF	Runic
tglg	1	U+	1700 ~	171F	Tagalog
hano	1	U+	1720 ~	173F	Hanunoo
buhd	1	U+	1740 ~	175F	Buhid
tagb	1	U+	1760 ~	177F	Tagbanwa
khmr	1	U+	1780 ~	17FF	Khmer
mong	1	U+	1800 ~	18AF	Mongolian
cans1	2	U+	18B0 ~	18FF	Unified Canadian Aboriginal Syllabics Extended
limb	1	U+	1900 ~	194F	Limbu
tale	1	U+	1950 ~	197F	Tai Le
talu	1	U+	1980 ~	19DF	New Tai Lue
khmr1	1	U+	19E0 ~	19FF	Khmer Symbols
bugi	1	U+	1A00 ~	1A1F	Buginese
lana	2	U+	1A20 ~	1AAF	Tai Tham
sym52	2	U+	1AB0 ~	1AFF	Combining Diacritical Marks Extended
bali	1	U+	1B00 ~	1B7F	Balinese
sund	2	U+	1B80 ~	1BBF	Sundanese
batk	2	U+	1BC0 ~	1BFF	Batak
lepc	2	U+	1C00 ~	1C4F	Lepcha
olck	2	U+	1C50 ~	1C7F	Ol Chiki
cyrlC	2	U+	1C80 ~	1C8F	Cyrillic Extended-C
geor2	4	U+	1C90 ~	1CBF	Georgian Extended
sund1	2	U+	1CC0 ~	1CCF	Sundanese Supplement
sym38	2	U+	1CD0 ~	1CFF	Vedic Extensions
latn4	1	U+	1D00 ~	1D7F	Phonetic Extensions
latn5	1	U+	1D80 ~	1DBF	Phonetic Extensions Supplement
sym03	1	U+	1DC0 ~	1DFF	Combining Diacritical Marks Supplement
latn3	1	U+	1E00 ~	1EFF	Latin Extended Additional
grek1	1	U+	1F00 ~	1FFF	Greek Extended
sym04	1	U+	2000 ~	206F	General Punctuation
sym05	1	U+	2070 ~	209F	Superscripts and Subscripts
sym06	1	U+	20A0 ~	20CF	Currency Symbols
sym07	1	U+	20D0 ~	20FF	Combining Diacritical Marks for Symbols
sym08	1	U+	2100 ~	214F	Letterlike Symbols
sym09	1	U+	2150 ~	218F	Number Forms
sym10	1	U+	2190 ~	21FF	Arrows
sym11	1	U+	2200 ~	22FF	Mathematical Operators
sym12	1	U+	2300 ~	23FF	Miscellaneous Technical
sym13	1	U+	2400 ~	243F	Control Pictures
sym14	1	U+	2440 ~	245F	Optical Character Recognition

sym15	1	U+	2460 ~	24FF	Enclosed Alphanumerics
sym16	1	U+	2500 ~	257F	Box Drawing
sym17	1	U+	2580 ~	259F	Block Elements
sym18	1	U+	25A0 ~	25FF	Geometric Shapes
sym19	1	U+	2600 ~	26FF	Miscellaneous Symbols
sym20	1	U+	2700 ~	27BF	Dingbats
sym21	1	U+	27C0 ~	27EF	Miscellaneous Mathematical Symbols-A
sym22	1	U+	27F0 ~	27FF	Supplemental Arrows-A
brai	1	U+	2800 ~	28FF	Braille Patterns
sym23	1	U+	2900 ~	297F	Supplemental Arrows-B
sym24	1	U+	2980 ~	29FF	Miscellaneous Mathematical Symbols-B
sym25	1	U+	2A00 ~	2AFF	Supplemental Mathematical Operators
sym26	1	U+	2B00 ~	2BFF	Miscellaneous Symbols and Arrows
glag	1	U+	2C00 ~	2C5F	Glagolitic
latnC	1	U+	2C60 ~	2C7F	Latin Extended-C
copt	1	U+	2C80 ~	2CFF	Coptic
geor1	1	U+	2D00 ~	2D2F	Georgian Supplement
tfng	1	U+	2D30 ~	2D7F	Tifinagh
ethi2	1	U+	2D80 ~	2DDF	Ethiopic Extended
cyrlA	2	U+	2DE0 ~	2DFF	Cyrillic Extended-A
sym27	1	U+	2E00 ~	2E7F	Supplemental Punctuation
cjk01	1	U+	2E80 ~	2EFF	CJK Radicals Supplement
cjk02	1	U+	2F00 ~	2FDF	Kangxi Radicals
cjk03	1	U+	2FF0 ~	2FFF	Ideographic Description Characters
cjk04	1	U+	3000 ~	303F	CJK Symbols and Punctuation
hira	1	U+	3040 ~	309F	Hiragana
kana	1	U+	30A0 ~	30FF	Katakana
bopo	1	U+	3100 ~	312F	Bopomofo
hang2	1	U+	3130 ~	318F	Hangul Compatibility Jamo
cjk05	1	U+	3190 ~	319F	Kanbun
bopo1	1	U+	31A0 ~	31BF	Bopomofo Extended
cjk06	1	U+	31C0 ~	31EF	CJK Strokes
kana1	1	U+	31F0 ~	31FF	Katakana Phonetic Extensions
cjk07	1	U+	3200 ~	32FF	Enclosed CJK Letters and Months
cjk08	1	U+	3300 ~	33FF	CJK Compatibility
haniA	1	U+	3400 ~	4DBF	CJK Unified Ideographs Extension A
sym28	1	U+	4DC0 ~	4DFF	Yijing Hexagram Symbols
hani	1	U+	4E00 ~	9FFF	CJK Unified Ideographs
yiii	1	U+	A000 ~	A48F	Yi Syllables
yiii1	1	U+	A490 ~	A4CF	Yi Radicals
lisu	2	U+	A4D0 ~	A4FF	Lisu
vaii	2	U+	A500 ~	A63F	Vai
cyrlB	2	U+	A640 ~	A69F	Cyrillic Extended-B
bamu	2	U+	A6A0 ~	A6FF	Bamum
sym29	1	U+	A700 ~	A71F	Modifier Tone Letters
latnD	1	U+	A720 ~	A7FF	Latin Extended-D
sylo	1	U+	A800 ~	A82F	Syloti Nagri
sym39	2	U+	A830 ~	A83F	Common Indic Number Forms
phag	1	U+	A840 ~	A87F	Phags-pa
saur	2	U+	A880 ~	A8DF	Saurashtra

deva1	2	U+	A8E0	A8FF	Devanagari Extended
kali	2	U+	A900	A92F	Kayah Li
rjng	2	U+	A930	A95F	Rejang
hangA	2	U+	A960	A97F	Hangul Jamo Extended-A
java	2	U+	A980	A9DF	Javanese
mymrB	2	U+	A9E0	A9FF	Myanmar Extended-B
cham	2	U+	AA00	AA5F	Cham
mymrA	2	U+	AA60	AA7F	Myanmar Extended-A
tavt	2	U+	AA80	AADF	Tai Viet
mtei1	2	U+	AAE0	AAFF	Meetei Mayek Extensions
ethiA	2	U+	AB00	AB2F	Ethiopic Extended-A
latnE	2	U+	AB30	AB6F	Latin Extended-E
cher1	2	U+	AB70	ABBF	Cherokee Supplement
mtei	2	U+	ABC0	ABFF	Meetei Mayek
hang	1	U+	AC00	D7AF	Hangul Syllables
hangB	2	U+	D7B0	D7FF	Hangul Jamo Extended-B
spc01	1	U+	D800	DB7F	High Surrogates
spc02	1	U+	DB80	DBFF	High Private Use Surrogates
spc03	1	U+	DC00	DFFF	Low Surrogates
spc04	1	U+	E000	F8FF	Private Use Area
hani1	1	U+	F900	FAFF	CJK Compatibility Ideographs
latn6	1	U+	FB00	FB4F	Alphabetic Presentation Forms
arab2	1	U+	FB50	FDFF	Arabic Presentation Forms-A
spc05	1	U+	FE00	FE0F	Variation Selectors
cjk09	1	U+	FE10	FE1F	Vertical Forms
sym30	1	U+	FE20	FE2F	Combining Half Marks
cjk10	1	U+	FE30	FE4F	CJK Compatibility Forms
cjk11	1	U+	FE50	FE6F	Small Form Variants
arab3	1	U+	FE70	FEFF	Arabic Presentation Forms-B
cjk12	1	U+	FF00	FFEF	Halfwidth and Fullwidth Forms
spc06	1	U+	FFF0	FFFF	Specials
linb	1	U+	10000	1007F	Linear B Syllabary
linb1	1	U+	10080	100FF	Linear B Ideograms
sym31	1	U+	10100	1013F	Aegean Numbers
grek2	1	U+	10140	1018F	Ancient Greek Numbers
sym40	2	U+	10190	101CF	Ancient Symbols
sym41	2	U+	101D0	101FF	Phaistos Disc
lyci	2	U+	10280	1029F	Lycian
cari	2	U+	102A0	102DF	Carian
copt1	2	U+	102E0	102FF	Coptic Epact Numbers
ital	1	U+	10300	1032F	Old Italic
goth	1	U+	10330	1034F	Gothic
perm	2	U+	10350	1037F	Old Permic
ugar	1	U+	10380	1039F	Ugaritic
xpeo	1	U+	103A0	103DF	Old Persian
dsrt	1	U+	10400	1044F	Deseret
shaw	1	U+	10450	1047F	Shavian
osma	1	U+	10480	104AF	Osmanya
osge	2	U+	104B0	104FF	Osage
elba	2	U+	10500	1052F	Elbasan

aghb	2	U+ 10530 ~ 1056F	Caucasian Albanian
lina	2	U+ 10600 ~ 1077F	Linear A
cprt	1	U+ 10800 ~ 1083F	Cypriot Syllabary
armi	2	U+ 10840 ~ 1085F	Imperial Aramaic
palm	2	U+ 10860 ~ 1087F	Palmyrene
nbat	2	U+ 10880 ~ 108AF	Nabataean
hatr	2	U+ 108E0 ~ 108FF	Hatran
phnx	1	U+ 10900 ~ 1091F	Phoenician
lydi	2	U+ 10920 ~ 1093F	Lydian
mero	2	U+ 10980 ~ 1099F	Meroitic Hieroglyphs
merc	2	U+ 109A0 ~ 109FF	Meroitic Cursive
khar	1	U+ 10A00 ~ 10A5F	Kharoshthi
sarb	2	U+ 10A60 ~ 10A7F	Old South Arabian
narb	2	U+ 10A80 ~ 10A9F	Old North Arabian
mani	2	U+ 10AC0 ~ 10AFF	Manichaean
avst	2	U+ 10B00 ~ 10B3F	Avestan
prti	2	U+ 10B40 ~ 10B5F	Inscriptional Parthian
phli	2	U+ 10B60 ~ 10B7F	Inscriptional Pahlavi
phlp	2	U+ 10B80 ~ 10BAF	Psalter Pahlavi
orkh	2	U+ 10C00 ~ 10C4F	Old Turkic
hung	2	U+ 10C80 ~ 10CFF	Old Hungarian
rohg	4	U+ 10D00 ~ 10D3F	Hanifi Rohingya
sym42	2	U+ 10E60 ~ 10E7F	Rumi Numeral Symbols
yezi	4	U+ 10E80 ~ 10EBF	Yezidi
sogo	4	U+ 10F00 ~ 10F2F	Old Sogdian
sogd	4	U+ 10F30 ~ 10F6F	Sogdian
chrs	4	U+ 10FB0 ~ 10FDF	Chorasmian
elym	4	U+ 10FE0 ~ 10FFF	Elymaic
brah	2	U+ 11000 ~ 1107F	Brahmi
kthi	2	U+ 11080 ~ 110CF	Kaithi
sora	2	U+ 110D0 ~ 110FF	Sora Sompeng
cakm	2	U+ 11100 ~ 1114F	Chakma
mahj	2	U+ 11150 ~ 1117F	Mahajani
shrd	2	U+ 11180 ~ 111DF	Sharada
sinh1	2	U+ 111E0 ~ 111FF	Sinhala Archaic Numbers
khoj	2	U+ 11200 ~ 1124F	Khojki
mult	2	U+ 11280 ~ 112AF	Multani
sind	2	U+ 112B0 ~ 112FF	Khudawadi
gran	2	U+ 11300 ~ 1137F	Grantha
newa	2	U+ 11400 ~ 1147F	Newa
tirh	2	U+ 11480 ~ 114DF	Tirhuta
sidd	2	U+ 11580 ~ 115FF	Siddham
modi	2	U+ 11600 ~ 1165F	Modi
mong1	2	U+ 11660 ~ 1167F	Mongolian Supplement
takr	2	U+ 11680 ~ 116CF	Takri
ahom	2	U+ 11700 ~ 1174F	Ahom
dogr	4	U+ 11800 ~ 1184F	Dogra
wara	2	U+ 118A0 ~ 118FF	Warang Citi
diak	4	U+ 11900 ~ 1195F	Dives Akuru
nand	4	U+ 119A0 ~ 119FF	Nandinagari

zanb	3	U+ 11A00 ~	11A4F	Zanabazar Square
soyo	3	U+ 11A50 ~	11AAF	Soyombo
pauc	2	U+ 11AC0 ~	11AFF	Pau Cin Hau
bhks	2	U+ 11C00 ~	11C6F	Bhaiksuki
marc	2	U+ 11C70 ~	11CBF	Marchen
gonm	3	U+ 11D00 ~	11D5F	Masaram Gondi
gong	4	U+ 11D60 ~	11DAF	Gunjala Gondi
maka	4	U+ 11EE0 ~	11EFF	Makasar
lisu1	4	U+ 11FB0 ~	11FBF	Lisu Supplement
taml1	4	U+ 11FC0 ~	11FFF	Tamil Supplement
xsux	1	U+ 12000 ~	123FF	Cuneiform
xsux1	1	U+ 12400 ~	1247F	Cuneiform Numbers and Punctuation
xsux2	2	U+ 12480 ~	1254F	Early Dynastic Cuneiform
egyp	2	U+ 13000 ~	1342F	Egyptian Hieroglyphs
egyp1	4	U+ 13430 ~	1343F	Egyptian Hieroglyph Format Controls
hluw	2	U+ 14400 ~	1467F	Anatolian Hieroglyphs
bamu1	2	U+ 16800 ~	16A3F	Bamum Supplement
mroo	2	U+ 16A40 ~	16A6F	Mro
bass	2	U+ 16AD0 ~	16AFF	Bassa Vah
hmng	2	U+ 16B00 ~	16B8F	Pahawh Hmong
medf	4	U+ 16E40 ~	16E9F	Medefaidrin
plrd	2	U+ 16F00 ~	16F9F	Miao
cjk14	2	U+ 16FE0 ~	16FFF	Ideographic Symbols and Punctuation
tang	2	U+ 17000 ~	187FF	Tangut
tang1	2	U+ 18800 ~	18AFF	Tangut Components
kits	4	U+ 18B00 ~	18CFF	Khitan Small Script
tang2	4	U+ 18D00 ~	18D7F	Tangut Supplement
kana2	2	U+ 1B000 ~	1B0FF	Kana Supplement
kanaA	3	U+ 1B100 ~	1B12F	Kana Extended-A
kana3	4	U+ 1B130 ~	1B16F	Small Kana Extension
nshu	3	U+ 1B170 ~	1B2FF	Nushu
dupl	2	U+ 1BC00 ~	1BC9F	Duployan
sym53	2	U+ 1BCA0 ~	1BCAF	Shorthand Format Controls
sym32	1	U+ 1D000 ~	1D0FF	Byzantine Musical Symbols
sym33	1	U+ 1D100 ~	1D1FF	Musical Symbols
sym34	1	U+ 1D200 ~	1D24F	Ancient Greek Musical Notation
sym58	4	U+ 1D2E0 ~	1D2FF	Mayan Numerals
sym35	1	U+ 1D300 ~	1D35F	Tai Xuan Jing Symbols
sym36	1	U+ 1D360 ~	1D37F	Counting Rod Numerals
sym37	1	U+ 1D400 ~	1D7FF	Mathematical Alphanumeric Symbols
sgnw	2	U+ 1D800 ~	1DAAF	Sutton SignWriting
glag1	2	U+ 1E000 ~	1E02F	Glagolitic Supplement
hmnp	4	U+ 1E100 ~	1E14F	Nyiakeng Puachue Hmong
wcho	4	U+ 1E2C0 ~	1E2FF	Wancho
mend	2	U+ 1E800 ~	1E8DF	Mende Kikakui
adlm	2	U+ 1E900 ~	1E95F	Adlam
sym59	4	U+ 1EC70 ~	1ECBF	Indic Siyaq Numbers
sym61	4	U+ 1ED00 ~	1ED4F	Ottoman Siyaq Numbers
sym51	2	U+ 1EE00 ~	1EEFF	Arabic Mathematical Alphabetic Symbols
sym43	2	U+ 1F000 ~	1F02F	Mahjong Tiles

sym44	2	U+ 1F030 〜 1F09F	Domino Tiles
sym46	2	U+ 1F0A0 〜 1F0FF	Playing Cards
sym45	2	U+ 1F100 〜 1F1FF	Enclosed Alphanumeric Supplement
cjk13	2	U+ 1F200 〜 1F2FF	Enclosed Ideographic Supplement
sym47	2	U+ 1F300 〜 1F5FF	Miscellaneous Symbols and Pictographs
sym48	2	U+ 1F600 〜 1F64F	Emoticons
sym54	2	U+ 1F650 〜 1F67F	Ornamental Dingbats
sym49	2	U+ 1F680 〜 1F6FF	Transport and Map Symbols
sym50	2	U+ 1F700 〜 1F77F	Alchemical Symbols
sym55	2	U+ 1F780 〜 1F7FF	Geometric Shapes Extended
sym56	2	U+ 1F800 〜 1F8FF	Supplemental Arrows-C
sym57	2	U+ 1F900 〜 1F9FF	Supplemental Symbols and Pictographs
sym60	4	U+ 1FA00 〜 1FA6F	Chess Symbols
sym62	4	U+ 1FA70 〜 1FAFF	Symbols and Pictographs Extended-A
sym63	4	U+ 1FB00 〜 1FBFF	Symbols for Legacy Computing
haniB	1	U+ 20000 〜 2A6DF	CJK Unified Ideographs Extension B
haniC	2	U+ 2A700 〜 2B73F	CJK Unified Ideographs Extension C
haniD	2	U+ 2B740 〜 2B81F	CJK Unified Ideographs Extension D
haniE	2	U+ 2B820 〜 2CEAF	CJK Unified Ideographs Extension E
haniF	3	U+ 2CEB0 〜 2EBEF	CJK Unified Ideographs Extension F
hani2	1	U+ 2F800 〜 2FA1F	CJK Compatibility Ideographs Supplement
haniG	4	U+ 30000 〜 3134F	CJK Unified Ideographs Extension G
spc07	1	U+ E0000 〜 E007F	Tags
spc08	1	U+ E0100 〜 E01EF	Variation Selectors Supplement
spc09	1	U+ F0000 〜 FFFFF	Supplementary Private Use Area-A
spc10	1	U+100000 〜 10FFFF	Supplementary Private Use Area-B

■ブロック ID の命名規則（参考）

- Unicode ブロック名にスクリプト（用字系）の名前が含まれるものは、それに対する ISO 15924 のコード（4 文字）を用いた。単一のスクリプトのブロックが複数ある場合は、名前が“Extended-A，B，……”のものは A，B，……を、それ以外のものは 1，2，……（一部符号値順でない）を末尾に付加した。

  “Arabic”→ arab ；“Latin Extended-C”→ latnC

  なお、“Hiragana”のコードは hira だが“Katakana”は kana であることに注意。
- それ以外は、CJK 関係（cjk）、特殊用途（spc）、それ以外（sym）の 3 種類に恣意的に分類して、後ろに 2 桁の番号を付けた（cjk12 など）。

5.2 どのブロックが使用できるか

以下では、各モード CCV においてサポートされるブロックの全容について説明する。

■モード CCV が 1 の場合 CCV 1 では Unicode ブロックの再分割が存在しないので、サポートされるブロックの全体は以下のようになる。

- 前掲の通常ブロックの表で「CCV」欄が 1 のブロック全て。

■モード CCV が 2 の場合　CCV 2 では `cjk12`“Halfwidth and Fullwidth Forms”のブロックが次の 3 つの下位ブロックに分割されている。

`cjk1a`	*Halfwidth and Fullwidth Forms/Other*
`cjk1b`	*Halfwidth and Fullwidth Forms/Latin*
`cjk1c`	*Halfwidth and Fullwidth Forms/Kana*

- `cjk1b`： ASCII 英数字の全角互換形。
  `U+FF10`〜`U+FF19`, `U+FF21`〜`U+FF3A`, `U+FF41`〜`U+FF5A`
- `cjk1c`： カタカナの半角互換形。句読点等の記号は含まない。
  `U+FF66`〜`U+FF6F`, `U+FF71`〜`U+FF9D`
- `cjk1a`： `cjk12` から `cjk1b`、`cjk1c` を除いた残り。

従って、サポートされるブロックの全体は以下のようになる。

- 前掲の通常ブロックの表で「CCV」欄が 2 以下のブロック全て。
  ※ `cjk12` のブロックも使用可能で、これは“Halfwidth and Fullwidth Forms”の全体を表す。
- `cjk12` の下位ブロック： `cjk1a`、`cjk1b`、`cjk1c` の 3 つ。

■モード CCV が 3 の場合　CCV 3 では `latn1`“Latin-1 Supplement”のブロックが次の 2 つの下位ブロックに分割されている。

`latnx`	*Latin-1 Supplement/Other*
`latny`	*Latin-1 Supplement/Latin*

- `latny`： ラテン文字であるもの。
  `U+FF10`〜`U+FF19`, `U+FF21`〜`U+FF3A`, `U+FF41`〜`U+FF5A`
- `latnx`： `latn1` から `latny` を除いた残り。

従って、サポートされるブロックの全体は以下のようになる。

- 前掲の通常ブロックの表にあるブロック全て。
  ※ `cjk12`、`latn1` のブロックも使用可能。
- `cjk12` の下位ブロック： `cjk1a`、`cjk1b`、`cjk1c` の 3 つ。
- `latn1` の下位ブロック： `latnx`、`latny` の 2 つ。

6　各モードにおける和文カテゴリの設定

（やっぱりあとで書く）