2015年3月18日
JGP会合(8)
資料4

日本語LGR-1、LGR-2概要(案)

ICANN52でのCJK合意を受けてのJGPの進め方(前回のおさらい)

- ICANN52でのCJK間調整会合において、次の3段階で調整を進めるというフレームワークを合意 - 以下はJGPを例にとり説明 -

step1

  - 日本語単独の観点で望ましいルール「日本語LGR-1」を作る
    - 中国語LGR-1、韓国語LGR-1のことは考えない(予想はするが、それを最初から受け入れた提案にはしない)

step2

  - CJK 3言語の言語LGR-1を統合するアルゴリズムを作る
  - そのアルゴリズムに従ってCJK統合後の日本語LGR-2を作る

step3

  - IPがCJK 3言語の言語LGR-2を統合しRootLGRを作る
    - IP自体がCJKの言語LGRに対して何らかの創造的作業をするわけでなく、各言語LRG-2をそのまま採用したルールの集合がRootLGRとなる

step1
日本語LGR-1を作る

- 日本語LGR-1で決めるべきこと ＝ 検討課題

a. 文字範囲(Repertoire)

b. 異体字(Variants)

c. 処置(Disposition) (*1)

d. ラベル評価ルール(Whole Label Evaluation（WLE) Rules) (*2)

(*1) 処置は、異体字が定義されている場合に個々の異体字の取り扱いを決めるもので、個々の文字に対して「登録可能」、「ブロック」などが定義される

(*2) 文字列(ラベル)全体の有効性を評価するルールで、例えば、中国語では簡体字と繁体字の混在は登録不可能とする、などが定義される

a. 文字範囲

- 代表的選択肢 (*1)

JGP 仮合意

① 常用漢字・人名用漢字の範囲(2998文字)

日本国内の地名や歴史的名称を表現するには不十分である。例えば奈良県橿原市の「橿」はこの範囲に含まれない。

② JIS X 0208:2012の第一水準・第二水準の範囲(6358文字)

常用漢字・人名用漢字を包含し、かつ日本国内の地名や歴史的名称を表記でき、JPドメイン名でも14年にわたる実績があり有用性が確認されている(文字の多寡について特に苦情や要望が生じていない)。

- ただし、常用漢字のうち4文字(「叱」「剝」「塡」「頬」)はここに含まれない

③ JIS X 0213:2012の第三水準・第四水準までの範囲(10053文字)

平均的な日本人は通常使用しておらず、また理解(読み書き)が困難。

④ IICORE(*2)の範囲(9810文字)

JIS X 0203:2012と同様に理解(読み書き)が困難。

(*1) 平仮名(83文字)、片仮名(86文字)、平仮名・片仮名に準ずる文字(5文字)はいずれの選択肢でも入るものとする

(*2) ISO/IEC 10646:2003/ Amendment 1:2005の一部として発行された国際標準で、漢字圏(CJK)全体で日常生活の用を満たす漢字集合

b. 異体字

JGP
仮合意

- 代表的選択肢

① 異体字を定義しない

ドメイン名文字列では固有名詞も使われる。日本語においては、特に固有名詞では異体字ととらえてもよさそうなもの(例:「国」と「國」)も含め、文字はすべて別文字として使う事例が散見される。また、「異体字なし」というルールは、JPドメイン名で14年にわたる実績がありIDNラベルでの有用性が確認されている(「異体字なし」に対して特に苦情や要望が生じていない)。

② 異体字を定義する

異体字を定義し、そこだけが違う文字列同士(例:「新日鉄」と「新日鐵」)を一括りに扱うことにより、ドメイン名の混同や混乱を避けられる可能性がある。しかし、ドメイン名というコンテキストでの適切な異体字定義を、権威ある根拠に基づいて得ることは困難である。また、TLDラベルは、オンライン即時登録でなく申請、審査を経て創設されるものであるため、混同や混乱が予想される文字列は、その審査段階で創設を拒絶されることになる。

c. 処置

(注) 異体字を定義しない場合は処置を定義する必要はない

- 代表的選択肢 - 異体字を定義する場合 -

① 「a. 文字範囲」に含まれるすべての文字は、それが異体字か否かに関わらず「登録可能」とする

固有名詞と一般名詞が組み合わされた文字列などでは、異体字同士が1つの名前の中に同時に存在すること(例：國學院大学)があり、異体字のすべてがドメイン名ラベル内で利用可能であるべきである。

② 個々の異体字について処置を決める

ドメイン名というコンテキストでの適切な異体字の処置を、権威ある根拠に基づいて得ることは困難である。

d. ラベル評価ルール

- 代表的選択肢

① 日本語として不自然な文字の並びを排除するルール(例:カナ長音や繰り返し文字は文字列の先頭に来ない、漢字の後ろにはカナ長音は来ない)を定義する

いくつかの禁止ルールを作り出し、それをあらかじめ禁止しておくことにより、文字列の適切さの審査の一部を自動化できる。しかし、ドメイン名というコンテキストでの適切なラベル評価ルールを、権威ある根拠に基づいて定義することは困難である。

② 日本語用のラベル評価ルールは定義しない

TLDラベルは、オンライン即時登録でなく申請、審査を経て創設されるものであるため、ラベル評価ルールが定義されなくても日本語として不適切な文字列は、その審査段階で創設を拒絶されることになる。

JGP 仮合意

step2
CJK 3言語のLGR-1を統合する アルゴリズムを作る

JGP 本日議論

- ポイント
  - CJKの各言語GPが合意できるものであること
  - CJKのLGR-1を入力としたとき、LGR-2の「b. 異体字」と「c. 処置」が自動的に決まるものであること
  - 日本語LGR-1の文字範囲に入っているすべての文字に対し、中国語LGR-1や韓国語LGR-1で異体字が定義された場合でも、日本語TLDでは元の文字も中国語LGR-1や韓国語LGR-1で異体字とみなされる文字も両方使用可能となるようにすること
    - 例えば、「国」と「國」が中国語LGR-1で異体字であった場合でも、日本語TLD「.国中」を申請した登録者は「.國中」も登録利用可能なアルゴリズムとすること

LGR-2作成アルゴリズム(案)

- JGPを例として説明 -

CP　: Code Point
VP　: Variant Point
VPs : Variant Points

JGP
本
日
議
論

手順 説明

(1)　各言語LGR-1を用意する

(2)　各言語LGR-1の申請可能文字(CPとする)をキーに、各CPの異体字リスト(VPsとする)の和集合を作り、マージテーブルを作成する

(3)　マージテーブルから、日本語LGR-1内の各CPとそれに紐付くVPsの組を抜き出す

(4)　個々のCPに対し、対応するVPsの中の個々の文字(VP)の処置を決める

a.　日本語LGR-1にてそのCPのVPsにVPが含まれている場合、日本語LGR-1で決めた処置をそのVPの処置とする

b.　日本語LGR-1にてそのCPのVPsにVPが含まれていない場合、そのVPが日本語LGR-1のCPに含まれていれば処置は割当可能とし、含まれていなければブロックとする

(5)　CPとVPsの対応付けを補完(相互参照が完全な状態に)する

a.　上記(4)-bで処置がブロックとされた文字をCPとみなし、マージテーブルからそのCPとVPsの組を抜き出す

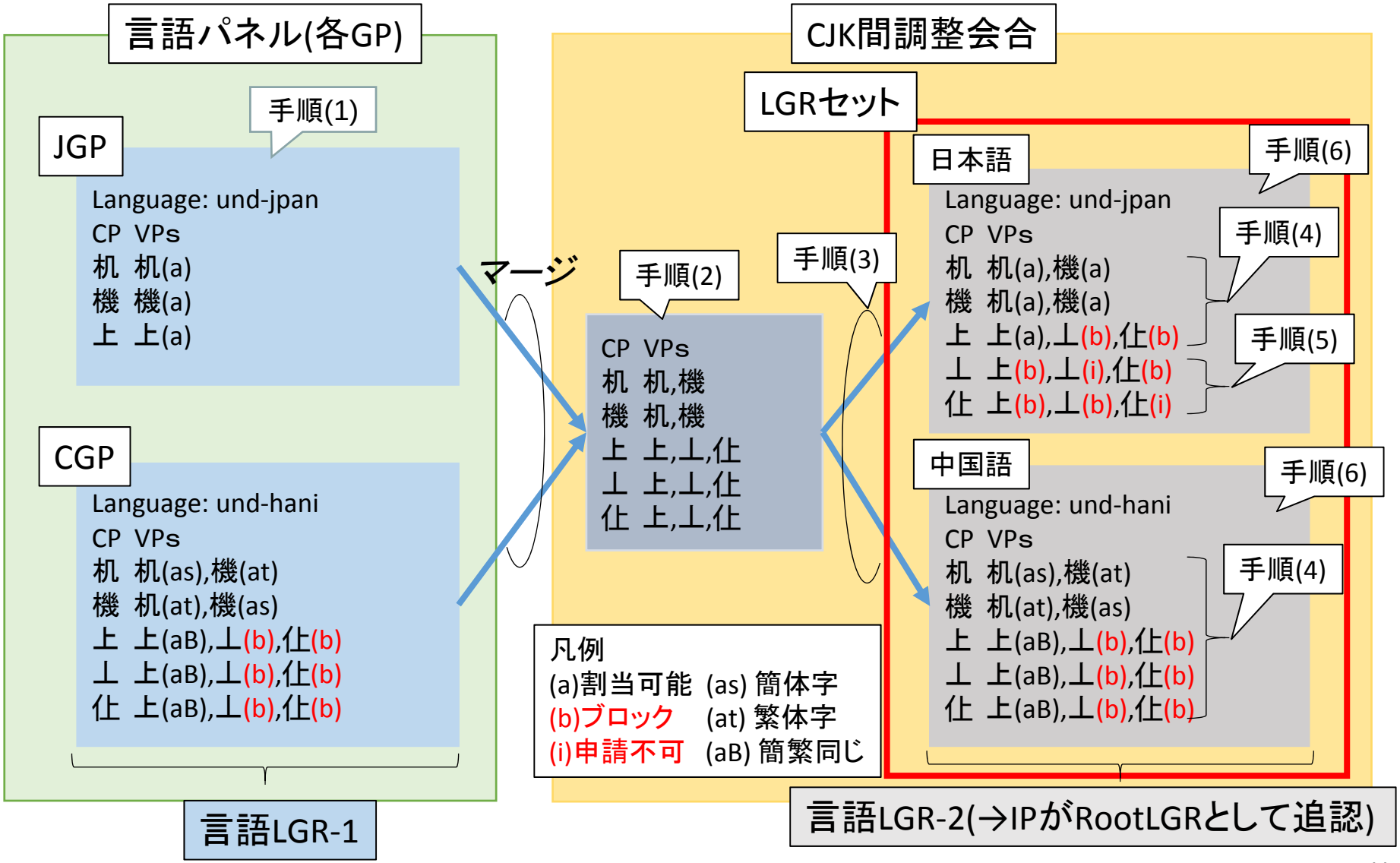
b.　そのCPのVPsにおける処置を申請不可能とし、その他のVPsのVPの処置をブロックとする

(6)　上記(5)と(6)の出力をマージしたものを統合後の日本語LGR(LGR-2)とする

【ケーススタディ】
「機上」についての考察

- 前提
  - 日本語LGR-1では文字範囲に「機」「机」「上」は入っているが「丄」と「仩」は入っていない
  - 日本語LGR-1では異体字が定義されない
  - 中国語LGR-1では「機」と「机」は異体字である
  - 中国語LGR-1では「上」と「丄」と「仩」は異体字である

アルゴリズムの適用例


言語パネル(各GP)
手順(1)
JGP
Language: und-jpan
CP VPs
机 机(a)
機 機(a)
上 上(a)
CGP
Language: und-hani
CP VPs
机 机(as),機(at)
機 机(at),機(as)
上 上(aB),丄(b),仕(b)
丄 上(aB),丄(b),仕(b)
仕 上(aB),丄(b),仕(b)
言語LGR-1
マージ
CJK間調整会合
LGRセット
手順(2)
CP VPs
机 机,機
機 机,機
上 上,丄,仕
丄 上,丄,仕
仕 上,丄,仕
手順(3)
日本語
手順(6)
Language: und-jpan
CP VPs
机 机(a),機(a)
機 机(a),機(a)
上 上(a),丄(b),仕(b)
丄 上(b),丄(i),仕(b)
仕 上(b),丄(b),仕(i)
手順(4)
手順(5)
中国語
手順(6)
Language: und-hani
CP VPs
机 机(as),機(at)
機 机(at),機(as)
上 上(aB),丄(b),仕(b)
丄 上(aB),丄(b),仕(b)
仕 上(aB),丄(b),仕(b)
手順(4)
凡例
(a)割当可能 (as) 簡体字
(b)ブロック (at) 繁体字
(i)申請不可 (aB) 簡繁同じ
言語LGR-2(→IPがRootLGRとして追認)

RootLGRの意味するところ

- 文字列を申請した場合の異体字含有文字列の扱い -

<日本語TLDとして申請した場合>

Language: und-jpan
申請文字列: 機上
割当可能: 機上,机上
ブロック: 機丄,機仩,机丄,机仩

Language: und-jpan
申請文字列: 机上
割当可能: 机上,機上
ブロック: 机丄,机仩,機丄,機仩

Language: und-jpan
申請文字列:机丄
(申請不可文字を含むため文字列は生成されない)

Language: und-jpan
申請文字列:機机
割当可能: 機机、機機、机机、机機
ブロック: (なし)

<中国語TLDとして申請した場合>

Language: und-hani
申請文字列: 機上
割当可能: 機上,机上
ブロック: 機丄,機仩,机丄,机仩

Language: und-hani
申請文字列: 机上
割当可能: 机上,機上
ブロック: 机丄,机仩,機丄,機仩

Language: und-hani
申請文字列: 机丄
割当可能: 机上,機上
ブロック: 机丄,机仩,機丄,機仩

Language: und-hani
申請文字列: 機机
割当可能: 機機、机机
ブロック: 機机、机機(簡繁混在は不可)