

Coordination Steps between CJK Generation Panels (Draft Proposal)

26th March 2015

Yoshiro Yoneya, JGP

Acknowledgement : Wang Wei, Ma Di, Yan Zhiwei

Precondition

- Each of CJK Generation Panels (GPs) generates LGR for each language TLD before integration
 - CJK GPs pick up ideographic variants for CJK from domain name usage perspective
 - CJK GPs don't elaborate ideographic variants for CJK from linguistic perspective
- CJK GPs agree on the mechanism (steps) to integrate and extract each language LGR

Step 1

- Each CJK GP generates its own LGR (hereinafter, LGR-1)
 - This step corresponds to Dr. Zhang's definition 1 as copied below (as of 2015-03-10)

1. Attempt to Redefine Variants Within One Language Context

VARIANT GROUP DEFINITION

Within one language context (~~CHS, CHT, KR, or JP~~ **Chinese, Japanese, or Korean**), the ideographs with different glyphs are defined as Variant Groups whose major /modern pronunciations, meanings and usages are the same.

Step 2 (1/2)

- CJK GPs collectively generate a merged table of each LGR-1 (hereinafter, LGR-M)
 - Repertoire (codepoints; CPs) of LGR-M is the union of all CJK LGR-1s
 - Variants of each CP in LGR-M is the union of variants (for the CP) defined in all CJK LGR-1s
 - LGR-M does not have
 - Language tag
 - Disposition of variants
 - WLE (Whole Label Evaluation rules)

Step 2 (2/2)

- This step corresponds to Dr. Zhang's definition 2 as copied below (as of 2015-03-10)

2. Attempt to Define Variants across Language Contexts

CJK VARIANT GROUP DEFINITION

In the CJK language environments, if in each Variant Group (~~CHS, CHT, KR, and JP~~ **Chinese, Japanese, or Korean**) there exists at least one identical ideograph, those Variants Groups shall be integrated as a CJK Variant Groups.

According to the ~~transitivity~~ **symmetricity and transitivity** principle, the members of Variant Groups in different language contexts can be treated as CJK Variants of each other.

~~In a CJK Variant Group, there will be one, two or three TYPICAL, ORTHOGRAPHIC, or PREFERRED VARIANTS coming from different languages.~~

Note: The last sentence is meaningless according to precondition

Step 3

- From LGR-M, each GP picks up the variants corresponding to every CP in its LGR-1
 - Disposition of variant(s) inherits its original disposition value in LGR-1
 - Disposition of variant(s) not defined in LGR-1 is defined as:
 - “blocked” if the variant is not in LGR-1 repertoire
 - “allocatable” otherwise (CJK consensus is needed on this)

Another option is : Otherwise, inherit its original disposition value in LGR-1 (one of “allocatable”, “simp”, “trad”, or “both”)

Step 4

- Each GP extracts additional CPs and corresponding variants from LGR-M to establish symmetricity
 - Additional CPs are the ones that are not CPs in LGR-1 but appear as VPs in LGR-M
 - Disposition of each additional CP is “out-of-repertoire-var”
 - Disposition of variant(s) of additional CPs are “blocked” if it is in LGR-1 repertoire, otherwise “out-of-repertoire-var”

Step 5

- Each GP appends WLE defined in LGR-1 to the result of Step4
- And add following WLE rules
 - `<action disp="invalid" any-variant="out-of-repertoire-var" />`

Step 6

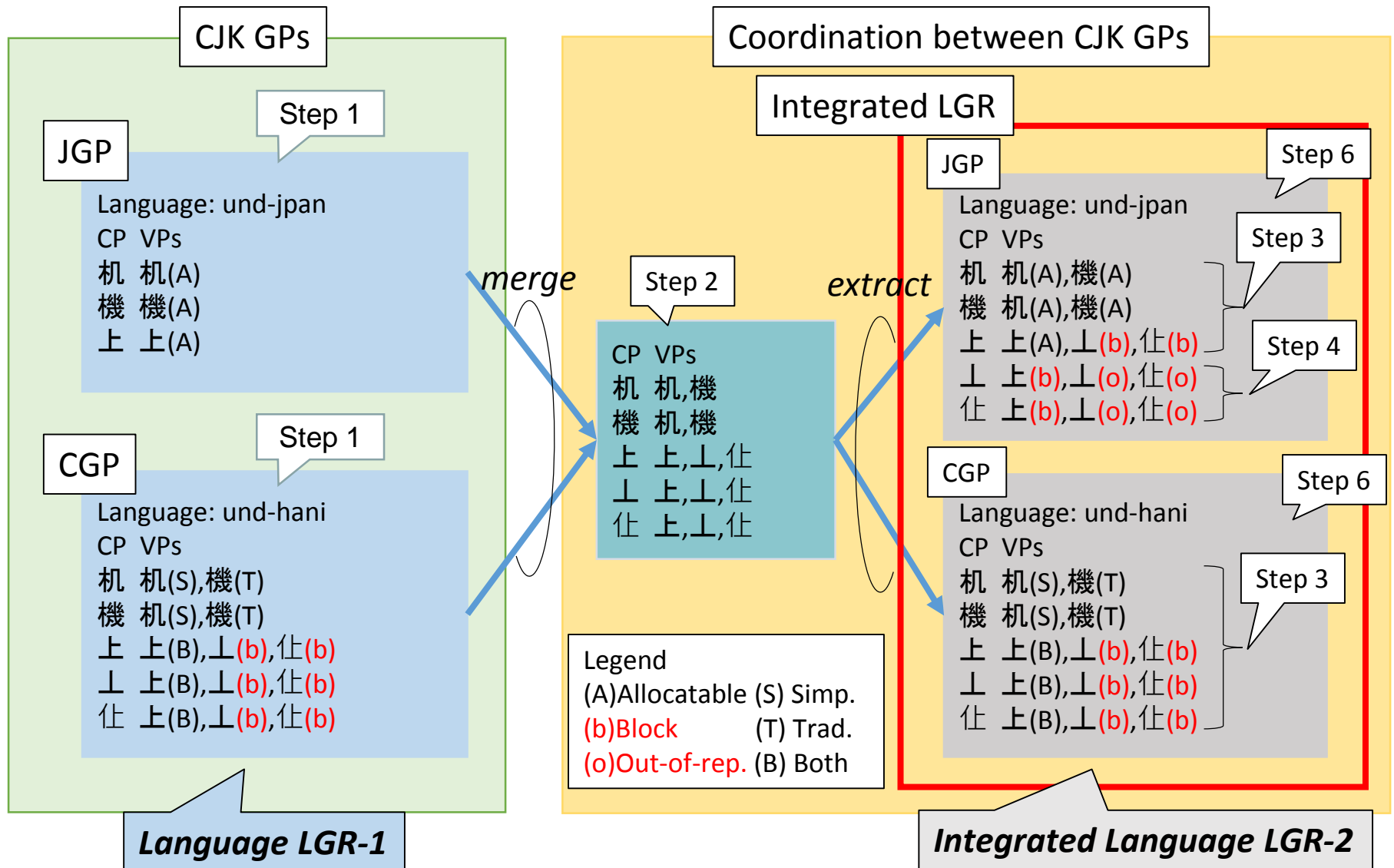
- Each GP compiles extracted CPs and its corresponding variants (with dispositions) and WLE into integrated language LGR (hereinafter, LGR-2)
 - And add language tag from its LGR-1

[Case Study]

Consideration on “機上”

- In case of JGP does not define any variants in Japanese LGR-1

Adoption of each step



Label generation examples

<In case of Japanese>

Language: und-jpan
Applied: 机上
Allocatable: 机上, 機上
blocked: 机丿, 机仕, 機丿, 機仕

Language: und-jpan
Applied: 机上
Allocatable: 机上, 機上
blocked: 机丿, 机仕, 機丿, 機仕

Language: und-jpan
Applied: 机丿
(no label generated due to out-of-repertoire-var)

Language: und-jpan
Applied: 機机
Allocatable: 机机, 机機, 機机, 機機
blocked: (nothing)

<In case of Chinese>

Language: und-hani
Applied: 机上
Allocatable: 机上, 機上
blocked: 机丿, 机仕, 機丿, 機仕

Language: und-hani
Applied: 机上
Allocatable: 机上, 機上
blocked: 机丿, 机仕, 機丿, 機仕

Language: und-hani
Applied: 机丿
Allocatable: 机上, 機上
blocked: 机丿, 机仕, 機丿, 機仕

Language: und-hani
Applied: 機机
Allocatable: 机机, 機機
blocked: 机機, 機机 (S/T mixed)

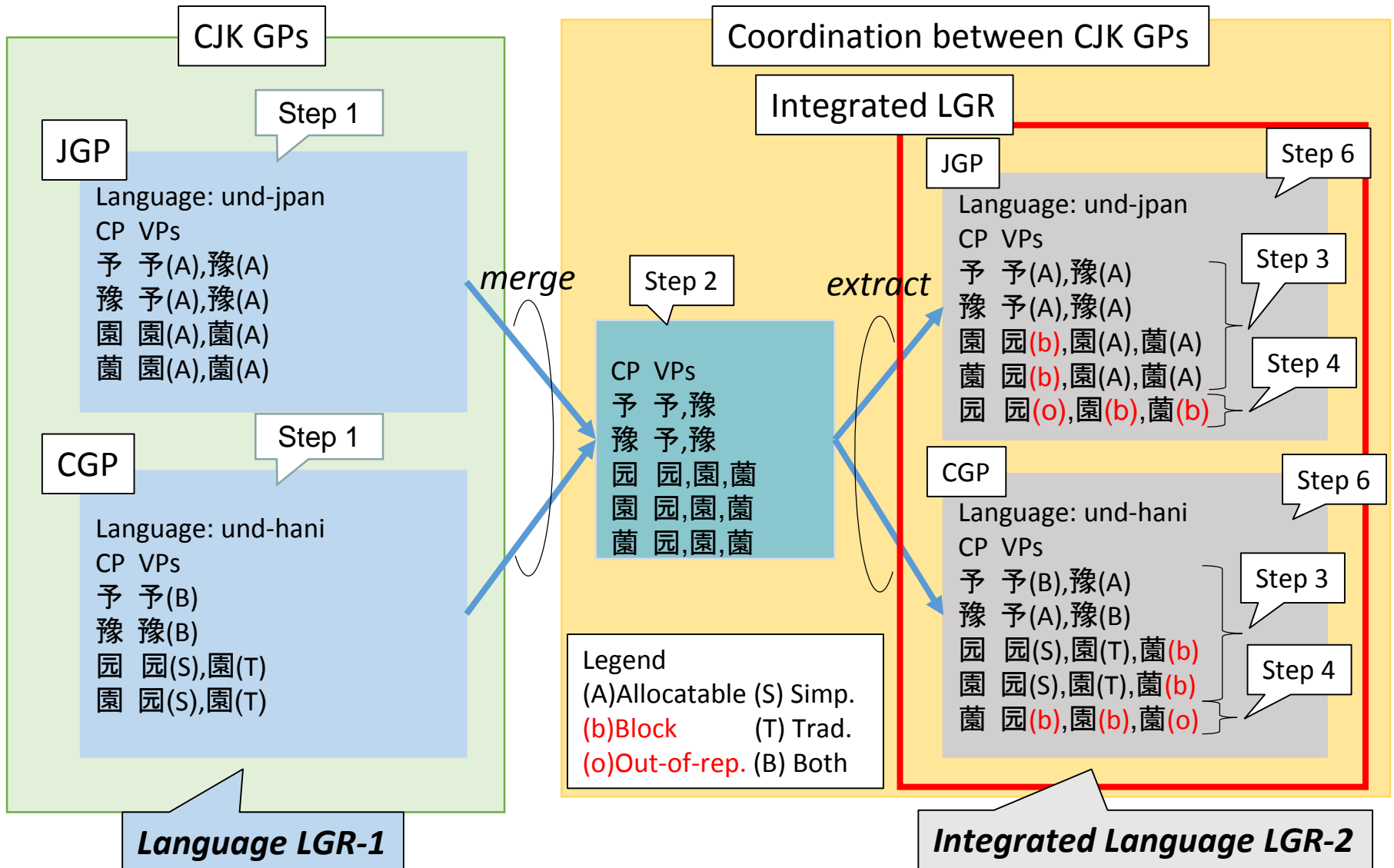
(blank page)

[Case Study]

Consideration on “豫園”

- In case of JGP defines variants in Japanese LGR-1

Adoption of each step



Label generation examples

<In case of Japanese>

Language: und-jpan
Applied: 豫園
Allocatable: 予園, 予菌, 豫園, 豫菌
blocked: 予园, 豫园

Language: und-jpan
Applied: 豫园
(no label generated due to
out-of-repertoire-var)

Language: und-jpan
Applied: 豫菌
Allocatable: 予園, 予菌, 豫園, 豫菌
blocked: 予园, 豫园

<In case of Chinese>

Language: und-hani
Applied: 豫園
Allocatable: 予园, 予園, 豫园, 豫園
blocked: 予菌, 豫菌

Language: und-hani
Applied: 豫园
Allocatable: 予园, 予園, 豫园, 豫園
blocked: 予菌, 豫菌

Language: und-hani
Applied: 豫菌
(no label generated due to
out-of-repertoire-var)