Ruling The Root

CJK Rules For The Root Zone

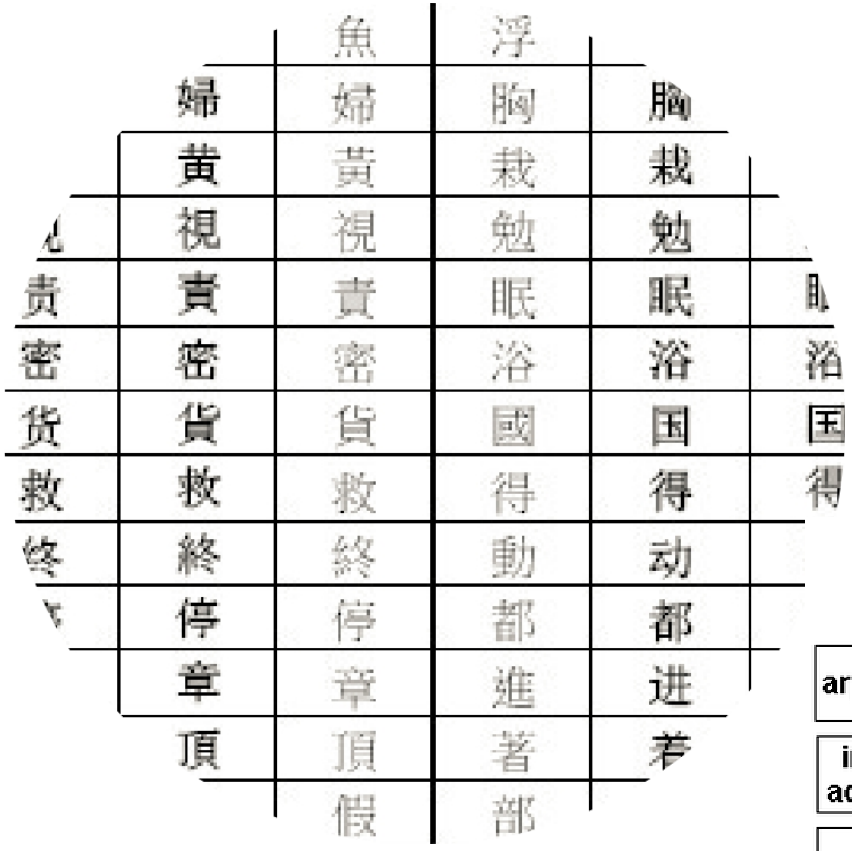
Kenny Huang, Ph.D. 黃勝雄博士

Member, CDNC / CGP
Co-author, RFC3743 IETF
Member, Executive Council, APNIC
Member, Board of Directors, TWNIC
huangksh@gmail.com

2014.Jun

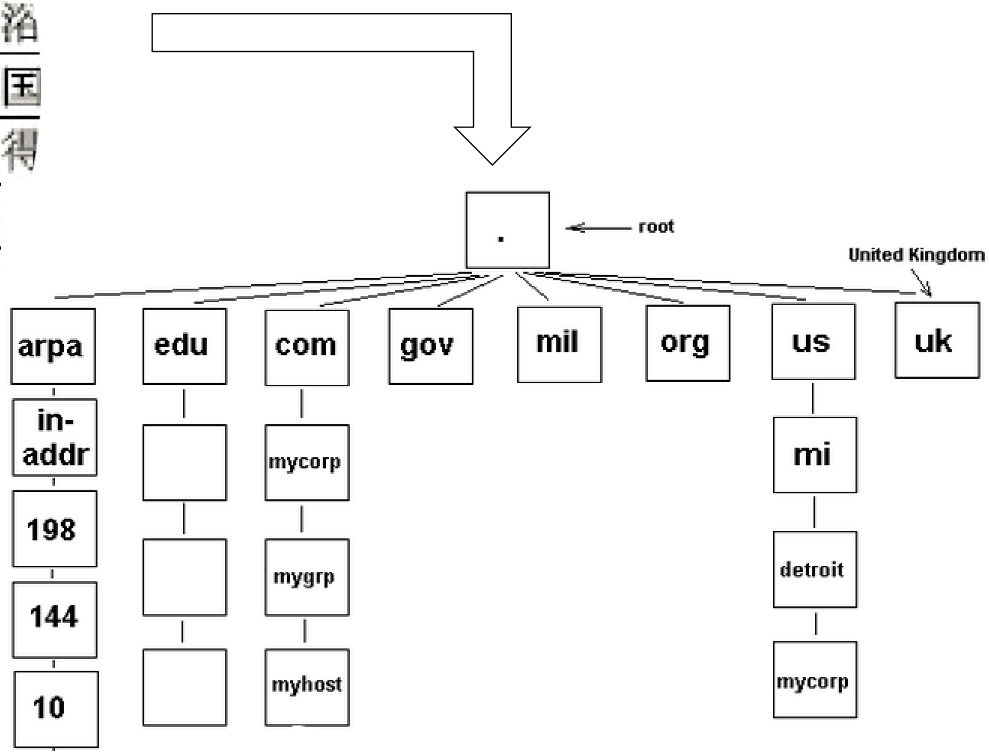
ICANNFIFTY
22-26 JUNE 2014

Problem : CJK Is Complicated

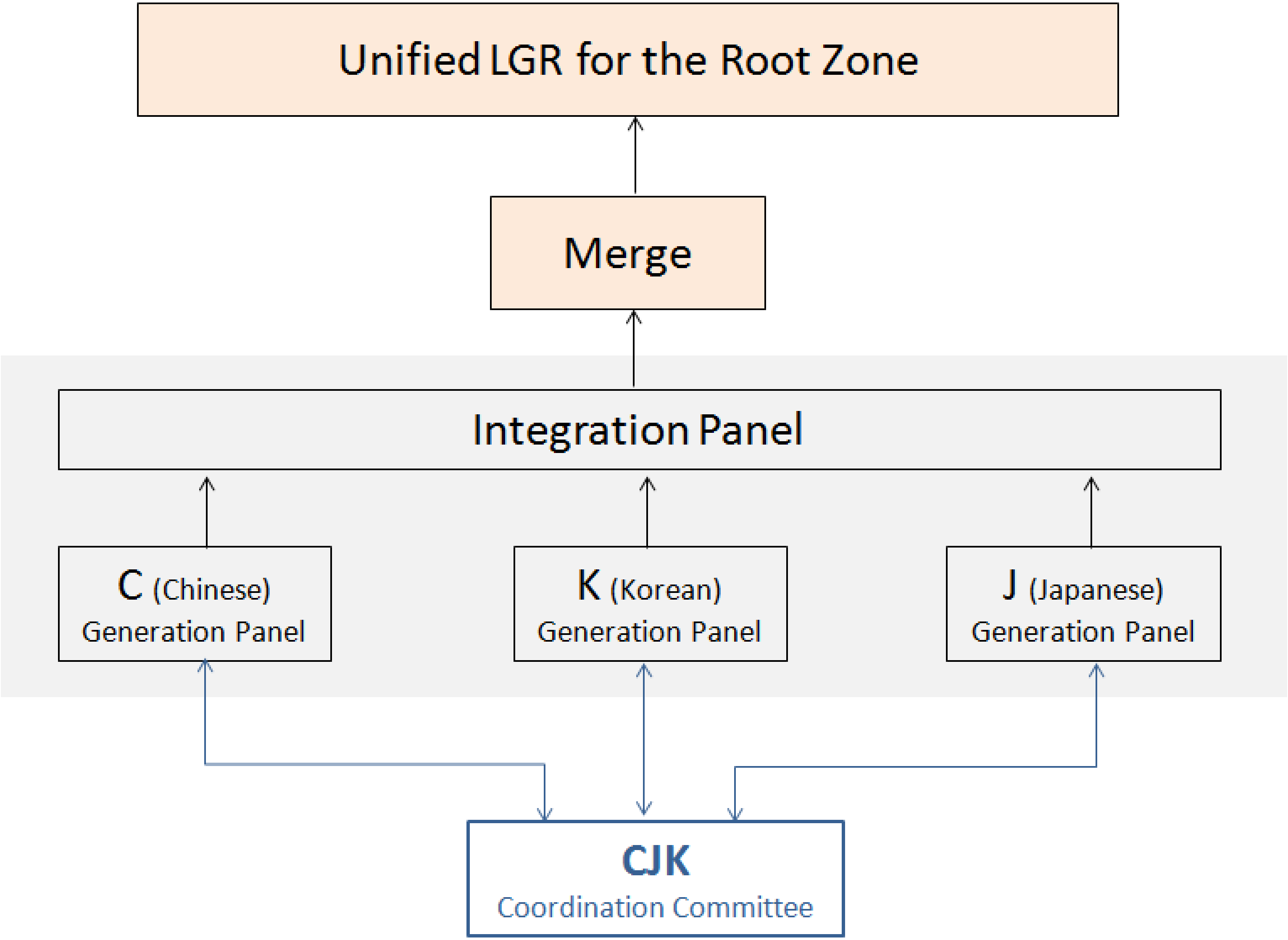


		魚	浮		
	婦	婦	胸	胸	
	黃	黃	栽	栽	
	視	視	勉	勉	
责	責	責	眠	眠	眠
密	密	密	浴	浴	浴
货	貨	貨	國	国	国
救	救	救	得	得	得
终	終	終	動	动	
	停	停	都	都	
	章	章	進	进	
	頂	頂	著	着	
		假	部		

Putting CJK labels in the root zone is even more complicated



Institutionalized Problem Solving : Structure



Unified LGR for the Root Zone

Merge

Integration Panel

C (Chinese)
Generation Panel

K (Korean)
Generation Panel

J (Japanese)
Generation Panel

CJK
Coordination Committee

Constraints for CJK LGR

Independent Tasks

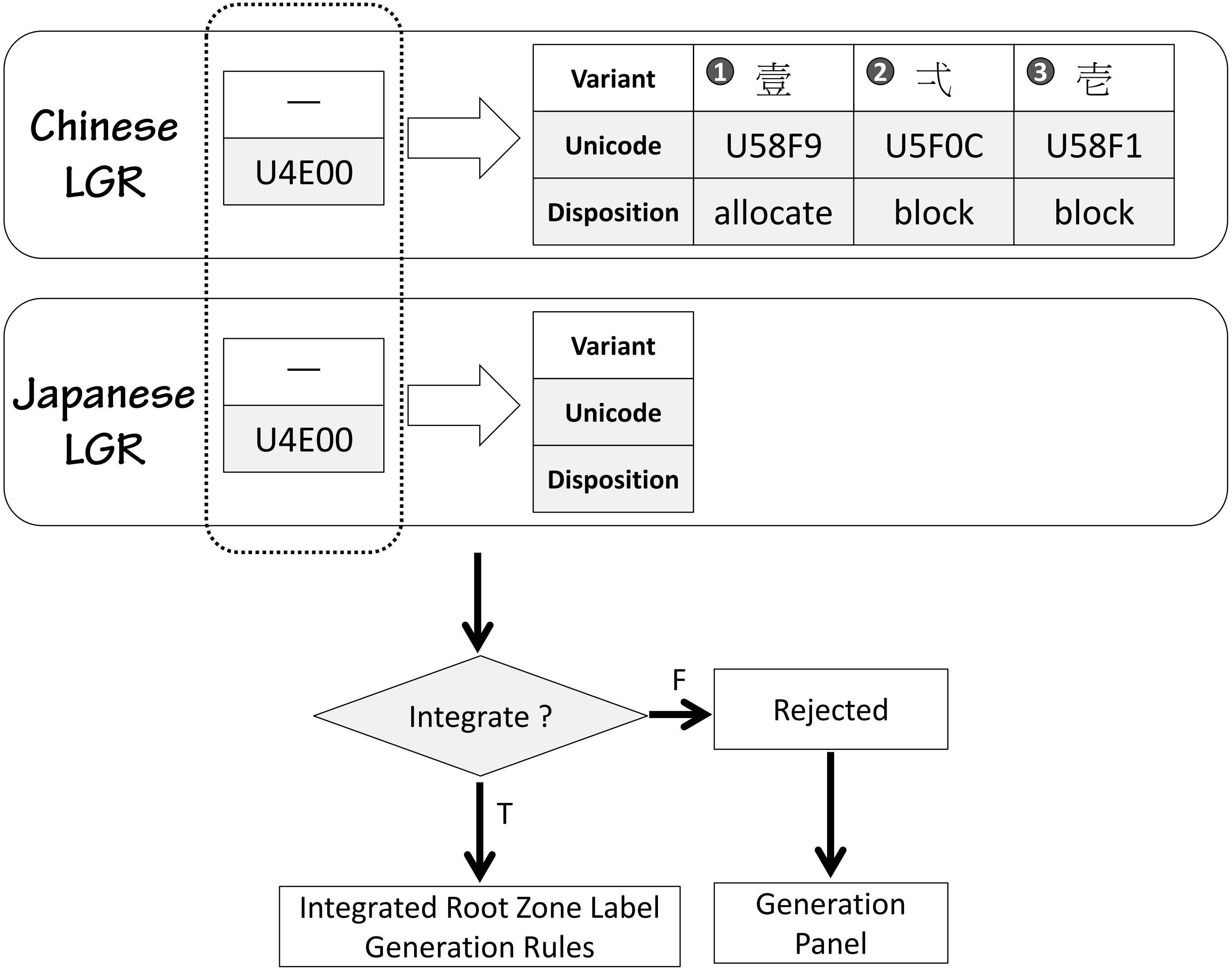
- ✓Each CJK Panel creates an LGR
- ✓Each LGR includes a repertoire and variants
  - ➢Define labels permission
  - ➢Define variants labels
  - ➢Assign dispositions
    - •Allocatable
    - •Block

Coordination Tasks

- ✓If an LGR includes Han characters:
  - ➢The variant *mappings* must agree for all the panels
  - ➢The variant *types* may be different
  - ➢The repertoires may be different

*Presented by Lee Han Chuan & IP, Shanghai 2014 May 29

Overlap Case Illustration



Chinese LGR

一
U4E00

Variant	❶ 壹	❷ 弌	❸ 壱
Unicode	U58F9	U5F0C	U58F1
Disposition	allocate	block	block

Japanese LGR

一
U4E00

Variant
Unicode
Disposition

Integrate ?

F

Rejected

T

Integrated Root Zone Label Generation Rules

Generation Panel

High Level Conflict Strategies

X

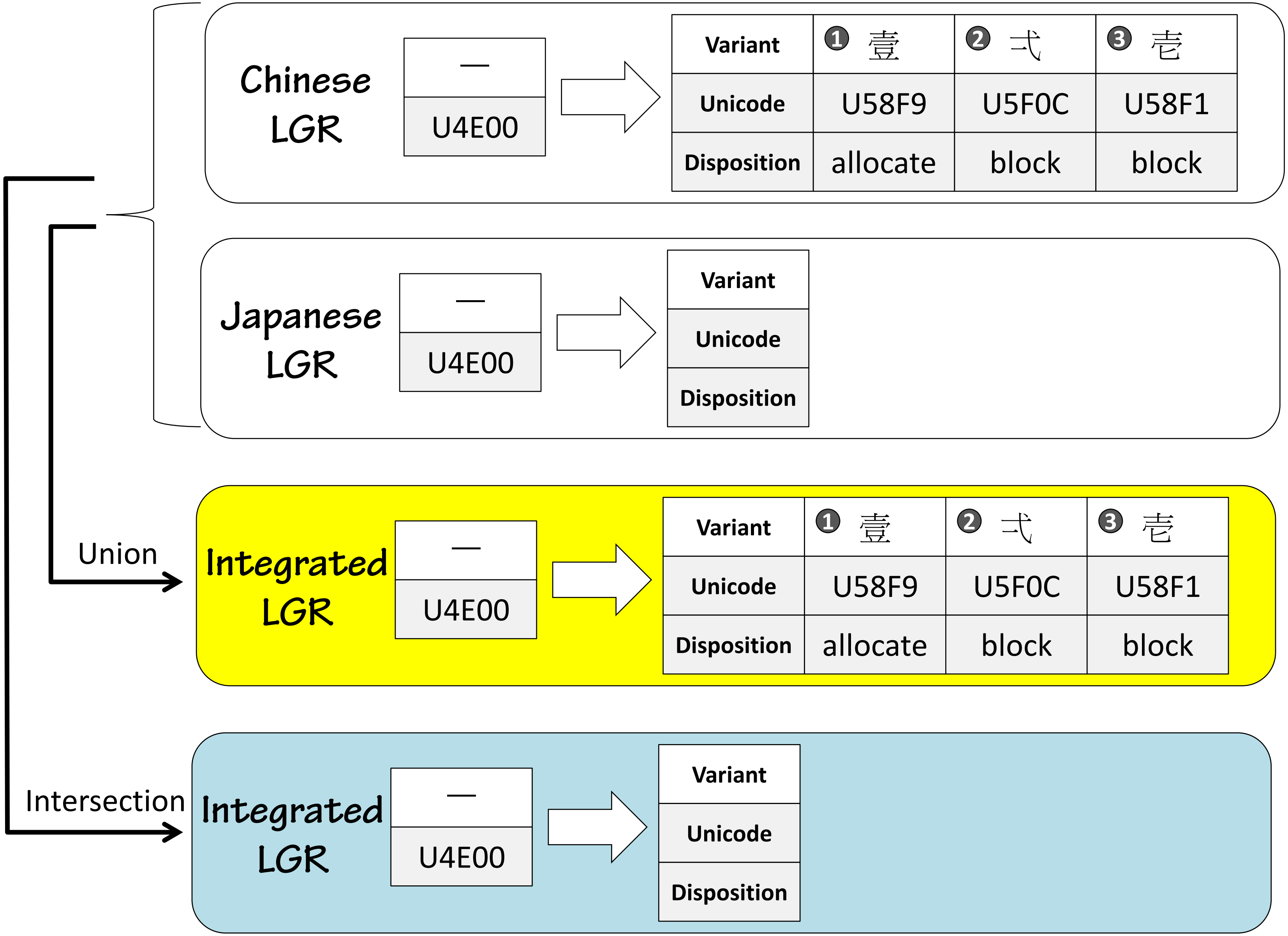
CJK overlap

C: rule R_c
J : rule R_j
K: rule R_k

R_{cjk} Conflict For Label X

ID	Strategy	Pros	Cons	Rank
1	Adopt X Abandon R_{cjk}	Permit X	No label rule	
2	Adopt X Intersection ∩ (R_{cjk})	Permit X Permit ∩(variants/disp)	Rules changed	
3	Adopt X Union ∪ (R_{cjk})	Permit X Permit ∪(variants/disp)	Rules changed	
4	Abandon X and R_{cjk}	No conflict	Label not available	
5	Adopt rules based on frequency of use	Fair & scientific approach	Rules changed; fairness doesn't mean appropriate	

Unified CJK LGR Illustration



Chinese LGR

一

U4E00

Variant	❶ 壹	❷ 弌	❸ 壱
Unicode	U58F9	U5F0C	U58F1
Disposition	allocate	block	block

Japanese LGR

一

U4E00

Variant
Unicode
Disposition

Union

Integrated LGR

一

U4E00

Variant	❶ 壹	❷ 弌	❸ 壱
Unicode	U58F9	U5F0C	U58F1
Disposition	allocate	block	block

Intersection

Integrated LGR

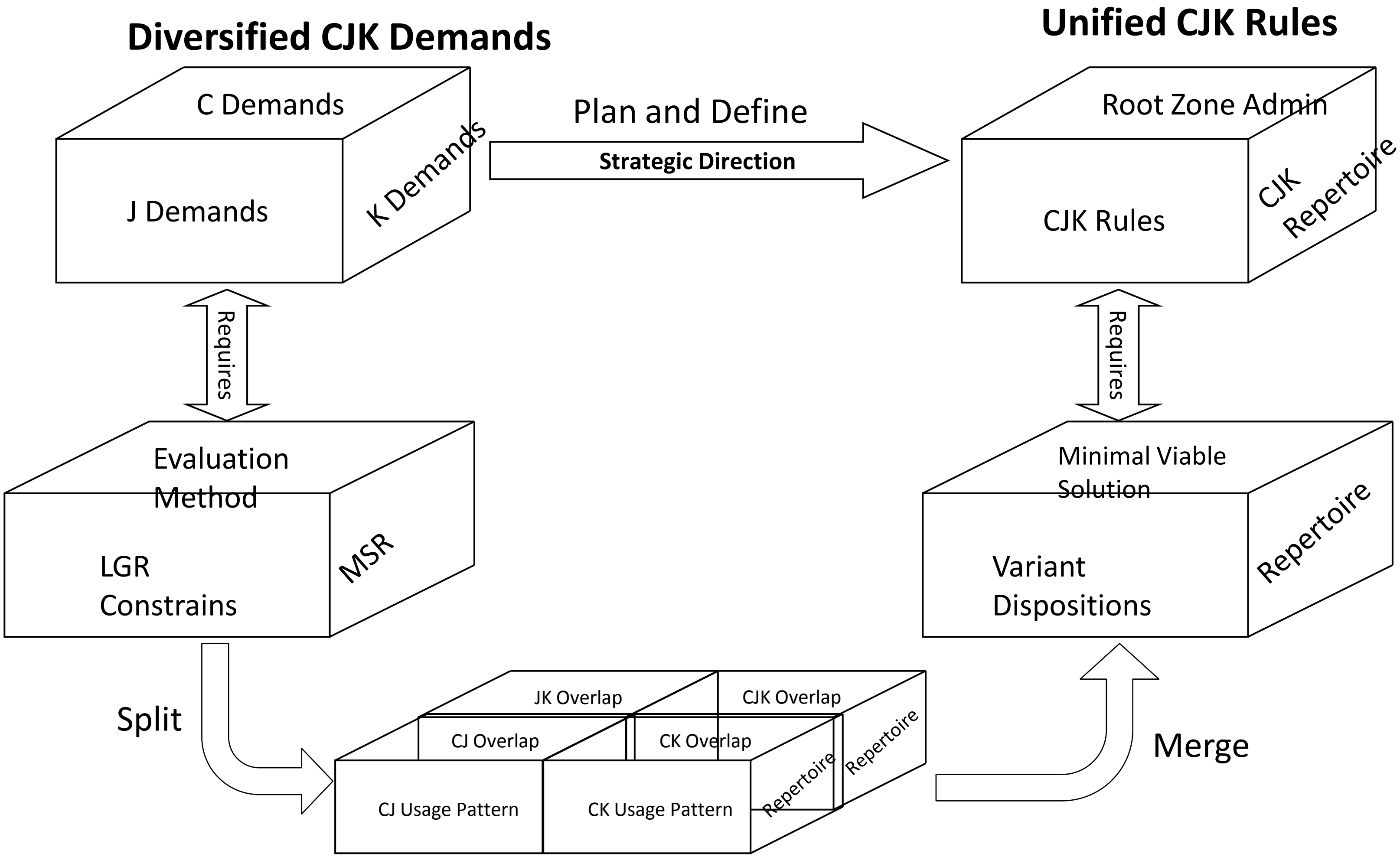
一

U4E00

Variant
Unicode
Disposition

CJK Integration Methodology

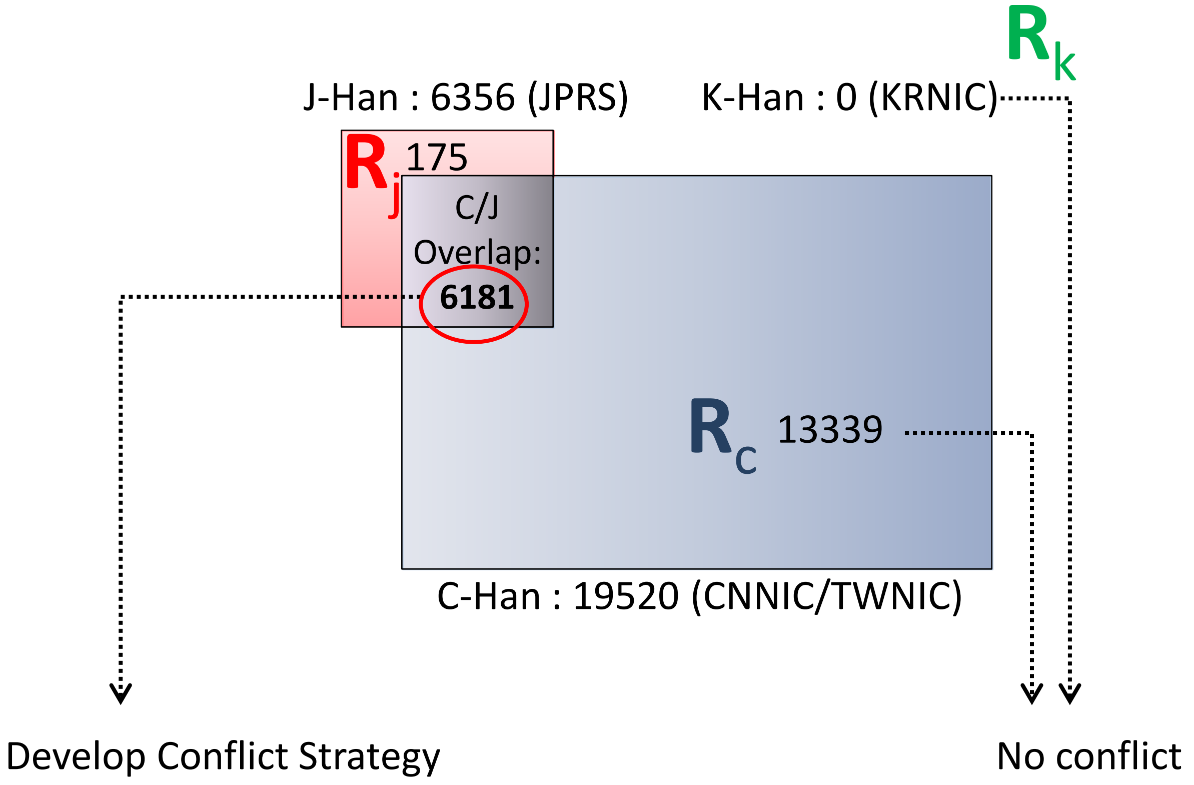
Divide & Conquer (D&C)



1

Splitting Non-overlapping Code Points From Repertories

CJK Han-overlap in IANA IDN Repository



R_k

J-Han : 6356 (JPRS)

K-Han : 0 (KRNIC)

R_j 175

C/J Overlap: **6181**

R_c 13339

C-Han : 19520 (CNNIC/TWNIC)

Develop Conflict Strategy

No conflict

R_c Chinese LGR

R_j Japanese LGR

R_k Korean LGR

unified code points
13339
+ 175
13514

Problem Domain (Unsolved Overlap) : 6181

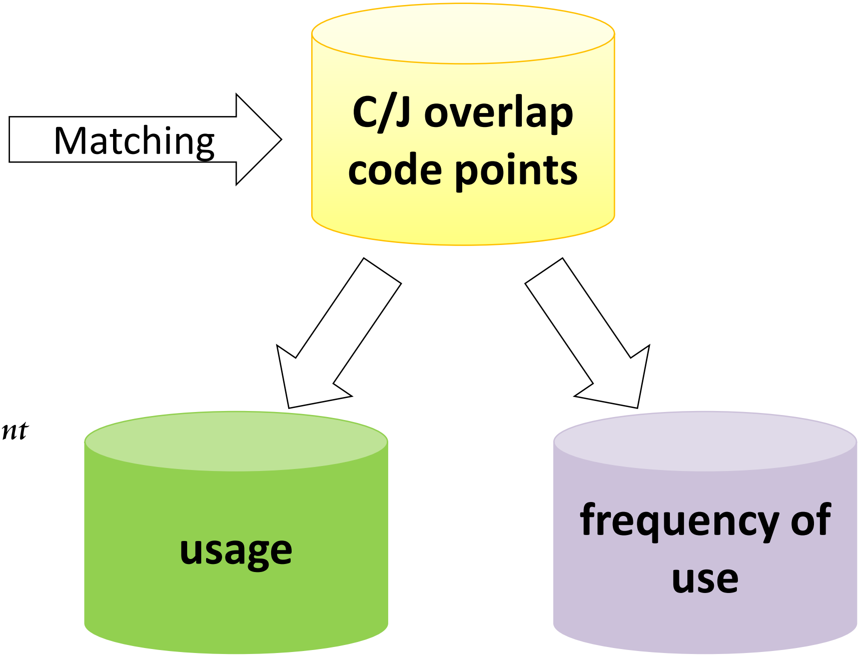
2

Engineering Design

Computation for Word Usage and Frequency

TC : Apple News
SC : Sina News
JP : Mainichi News

Sample size is statistical significant



Matching

C/J overlap code points

usage

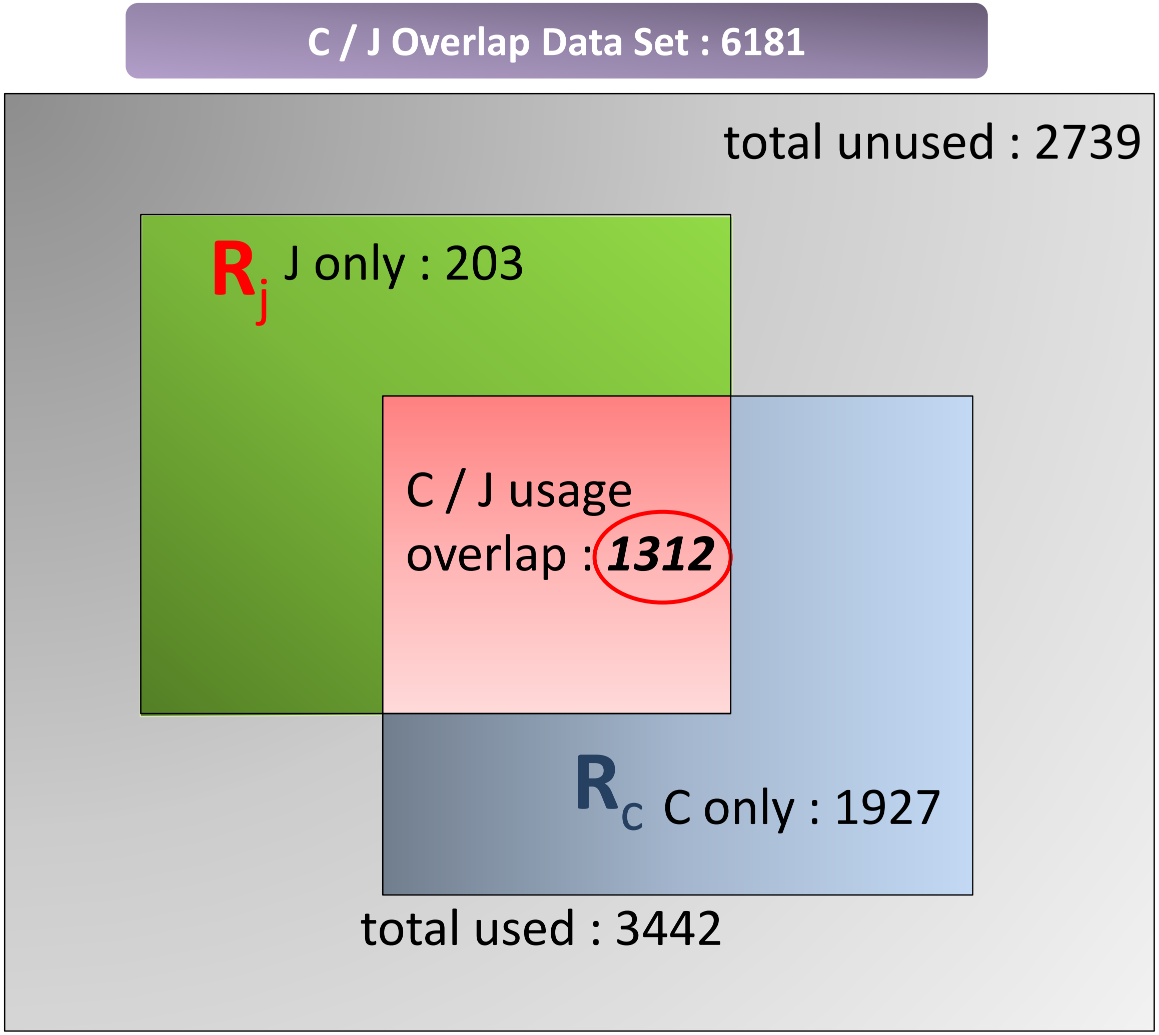
frequency of use

Split unused code points

Split code points of low frequency of use

3

Splitting Unused Code Points from The Overlap



C / J Overlap Data Set : 6181

total unused : 2739

R_j J only : 203

C / J usage overlap : ***1312***

R_C C only : 1927

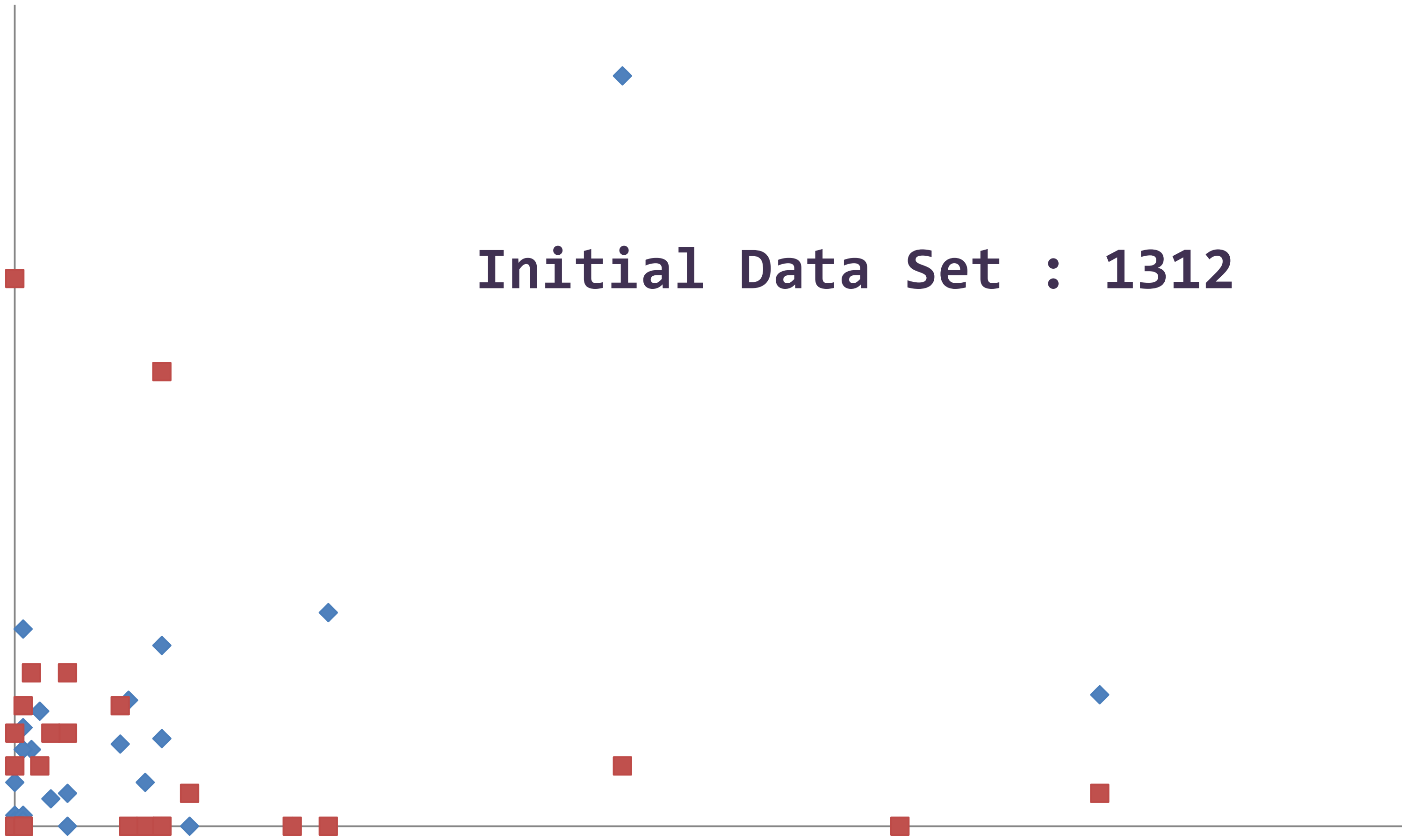
total used : 3442

unified code points
2739
203
+ 1927
4869

Problem Domain (Unsolved Overlap) : 1312

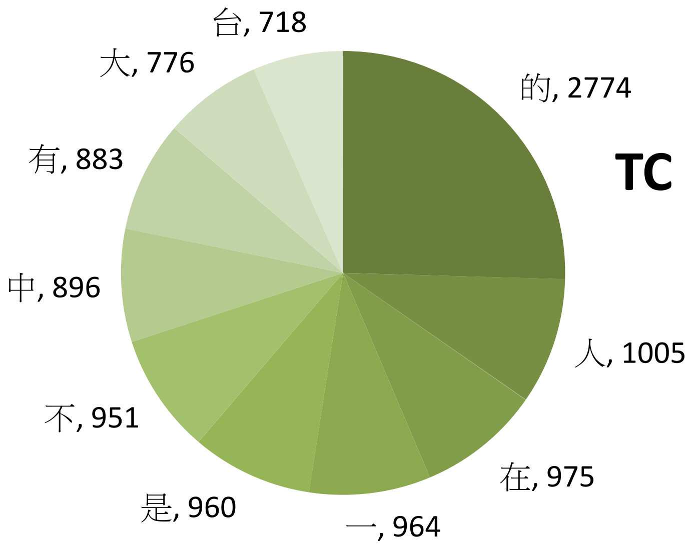
4

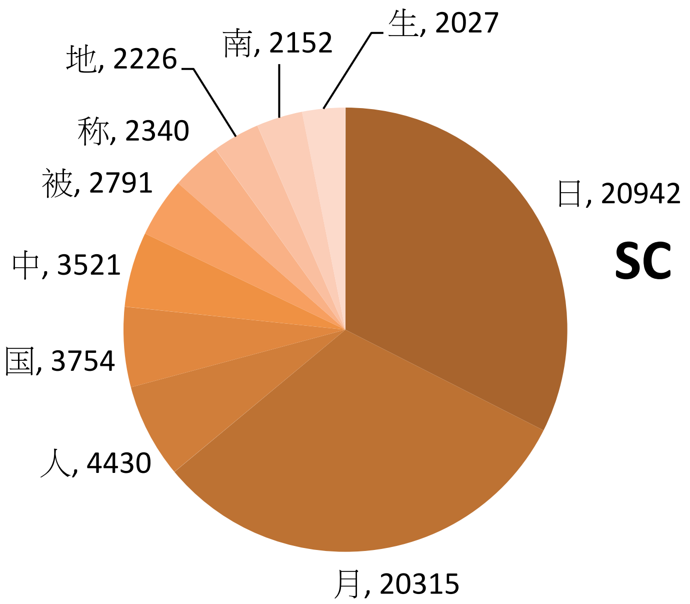
Computing Frequency of Use of Code Points

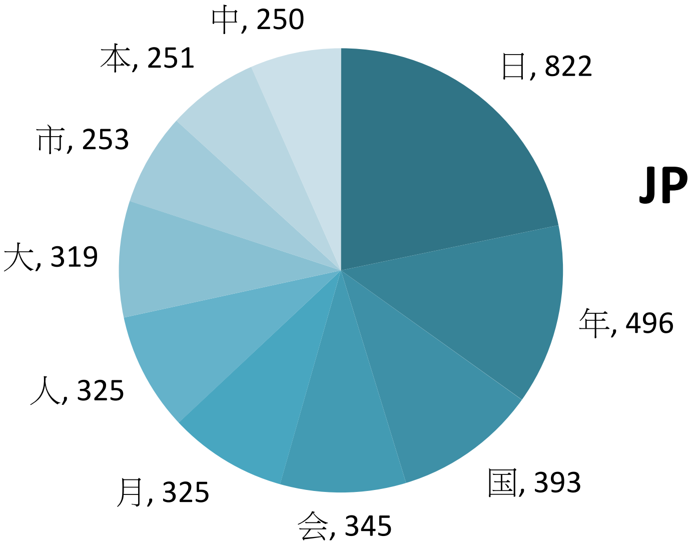


Initial Data Set : 1312

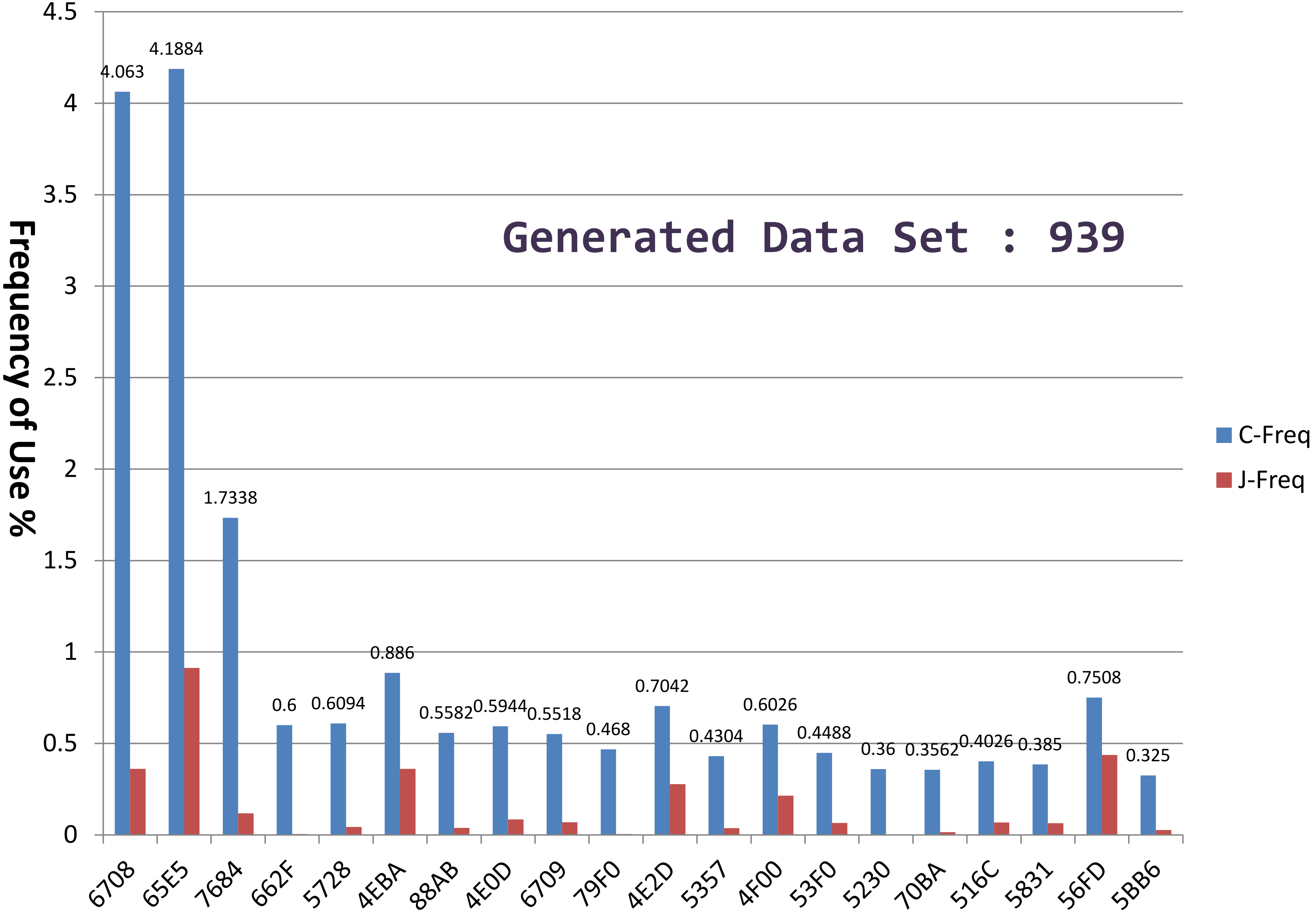
Top 10 Most Popular Words



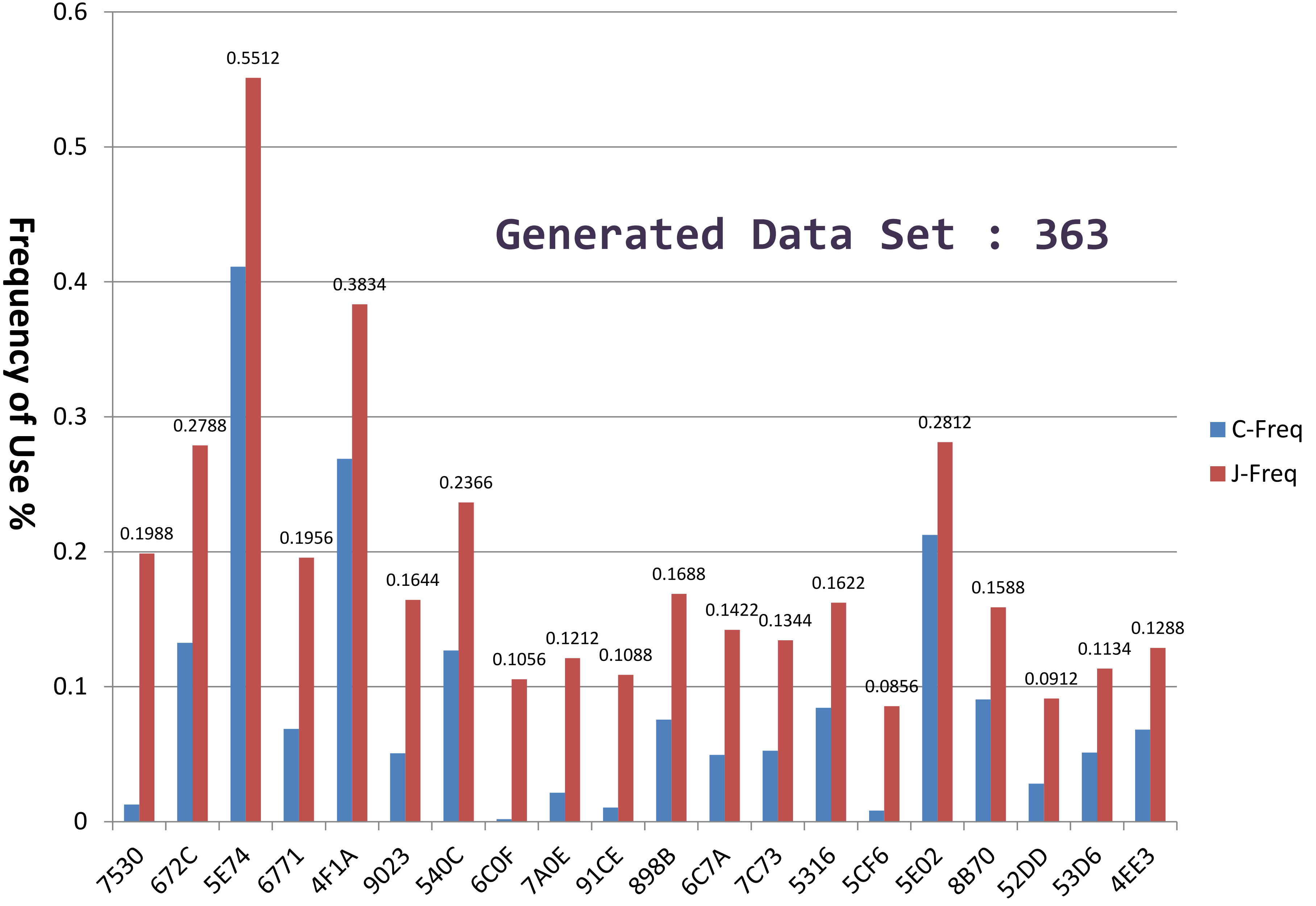




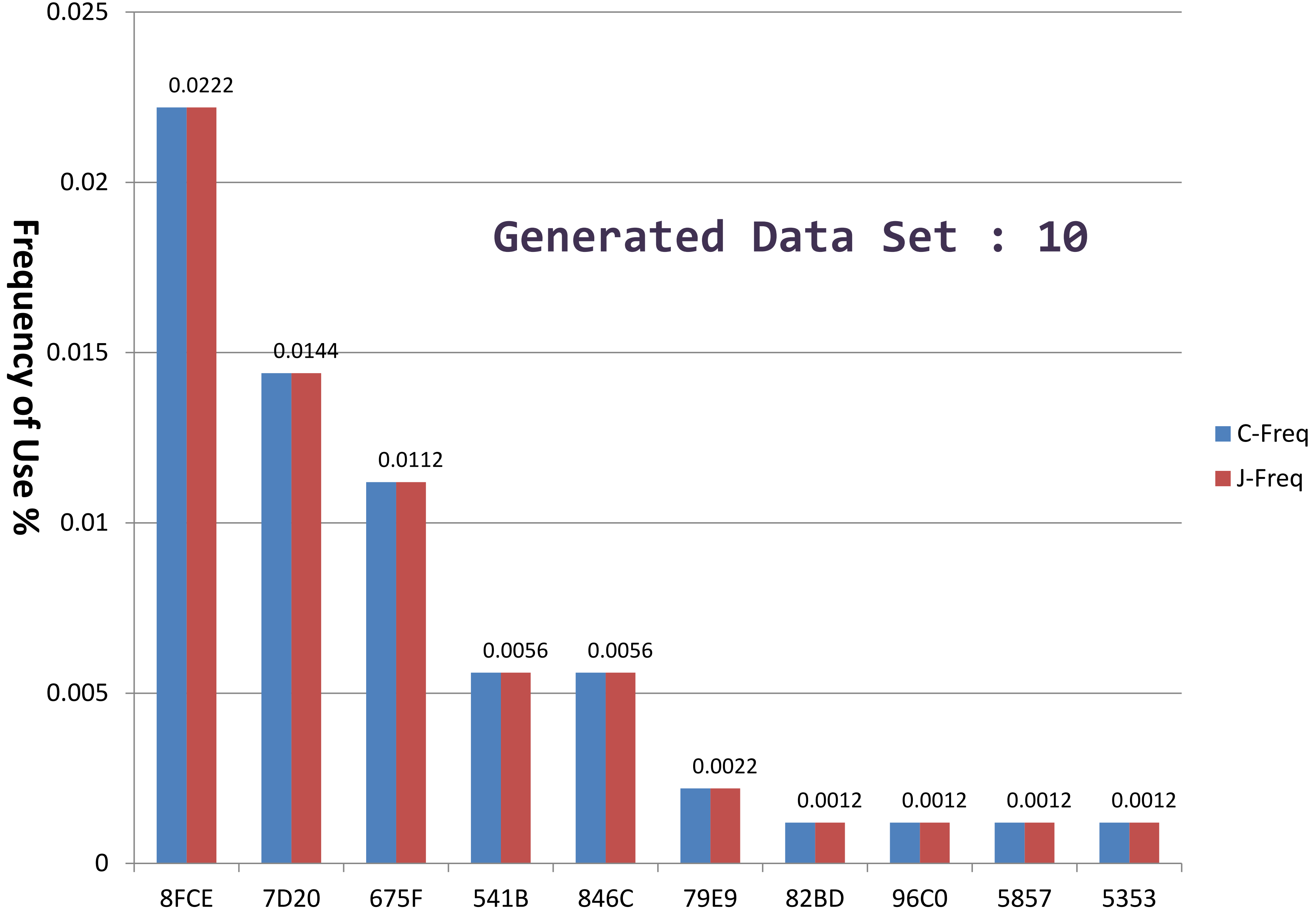
Top 20 : Chinese Frequency of Use > Japanese Frequency of Use



Top 20 : Chinese Frequency of Use < Japanese Frequency of Use

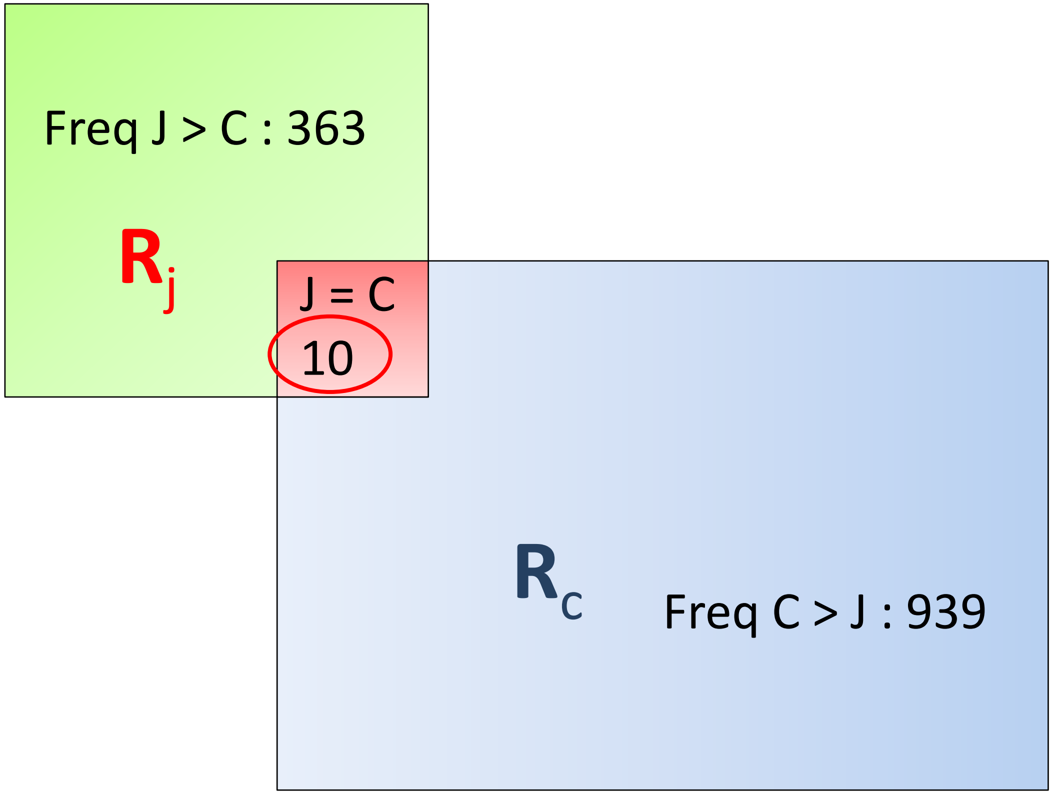


Chinese Frequency of Use = Japanese Frequency of Use



Frequency of Use Reassembly

C / J Usage Overlap Data Set : 1312



Freq J > C : 363

R_j

J = C

10

R_C

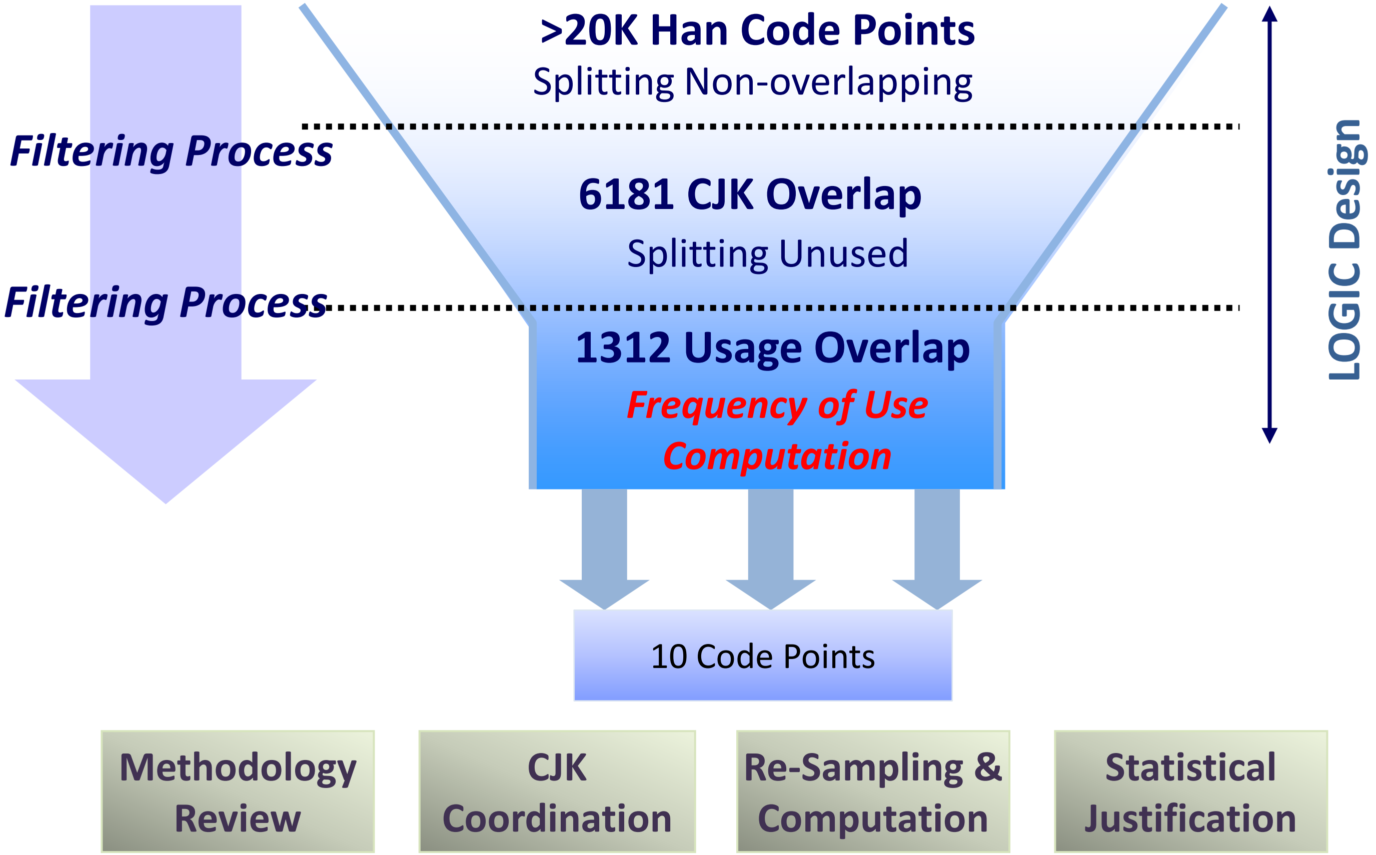
Freq C > J : 939

unified code points

363
+ 939
1302

Problem Domain (Unsolved Overlap) : 10

Data Processing & Computation Recap



>20K Han Code Points

Splitting Non-overlapping

Filtering Process

6181 CJK Overlap

Splitting Unused

Filtering Process

1312 Usage Overlap

Frequency of Use Computation

LOGIC Design

10 Code Points

Methodology Review

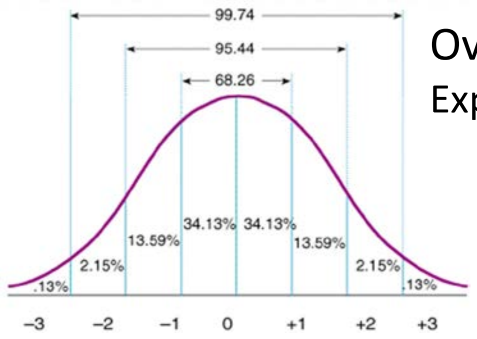
CJK Coordination

Re-Sampling & Computation

Statistical Justification

Problem domain was effectively reduced

Future Work



Overlap range redefine

Expand (?) Std Dev.

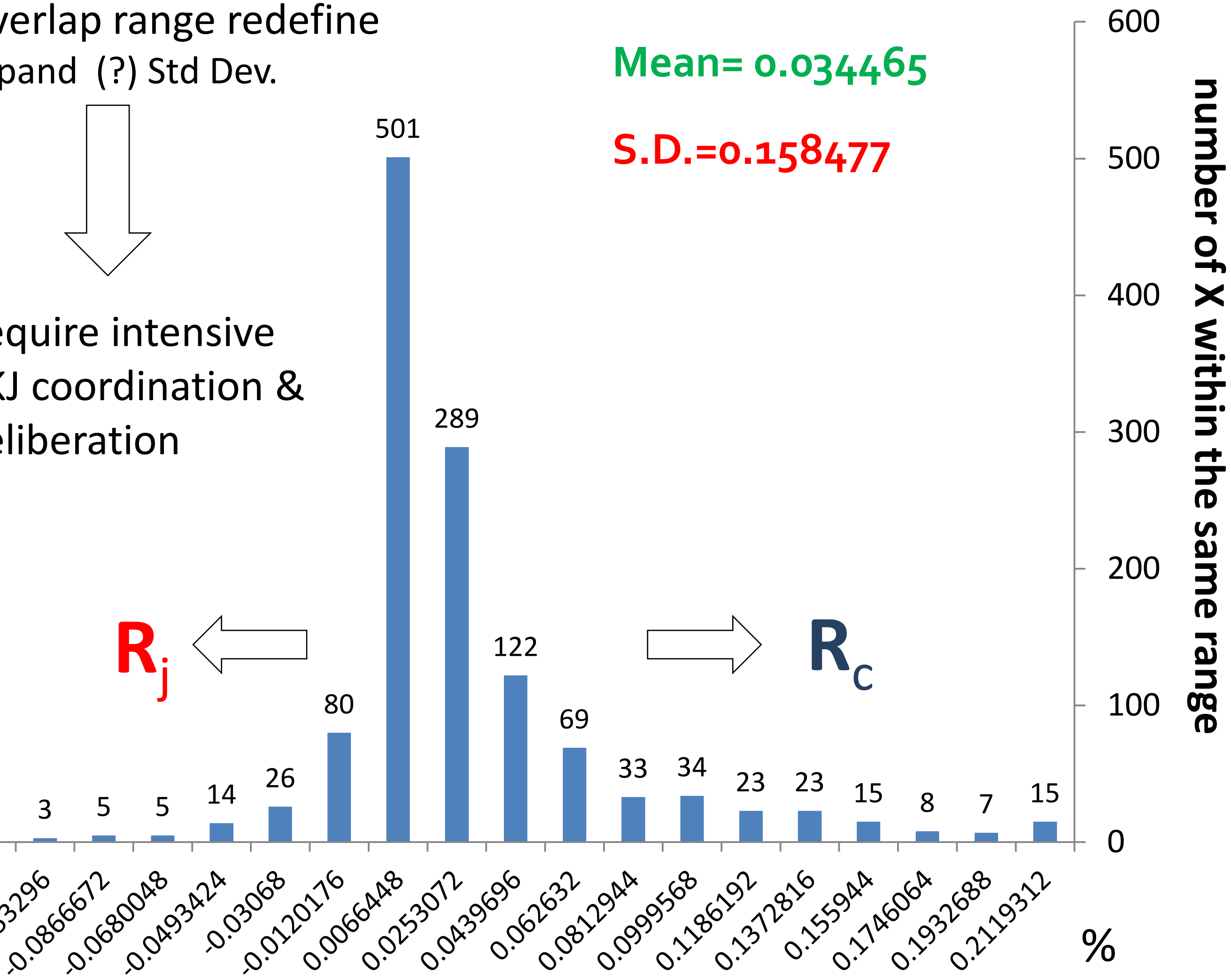
Require intensive CKJ coordination & deliberation

Mean= 0.034465

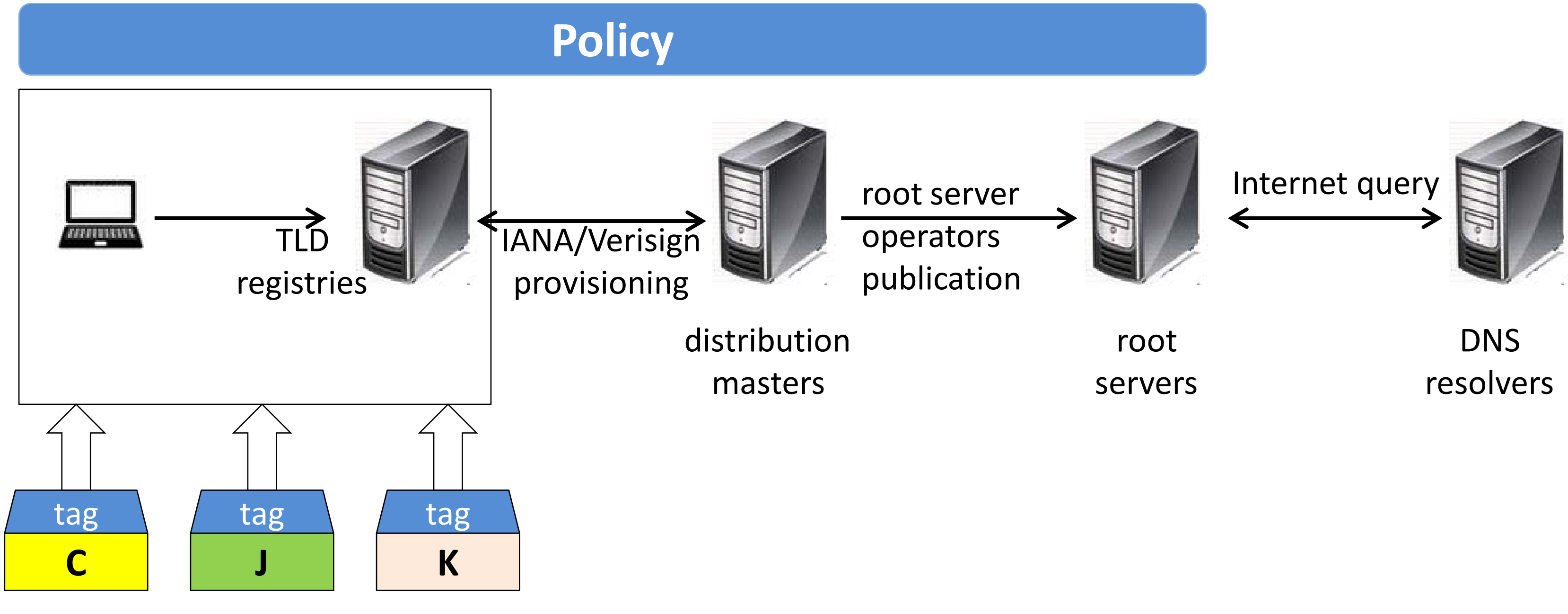
S.D.=0.158477

R_j

R_c



Re-consider Language Tag



Language tag support

Sources of Language Tag

- RFC 2860 : The name space of language tags is administered by IANA
- ISO Standard 639 :
  - when a language has both an IANA-registered tag and a tag derived from an ISO registered code, one MUST use the ISO tag.
  - Maintenance Agency : International Information Centre for Terminology (Austria)

Perfection Syndrome

"Engineering isn't about perfect solutions; it's about doing the best you can with limited resources." Randy Pausch

Thank You

Question?

Kenny Huang, Ph.D.
huangksh@gmail.com