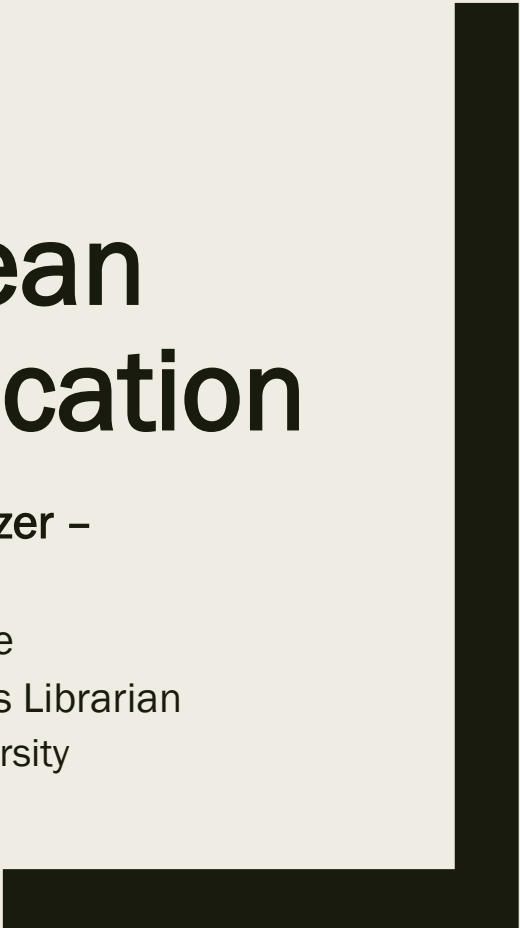


# Developing Korean Romanization Application

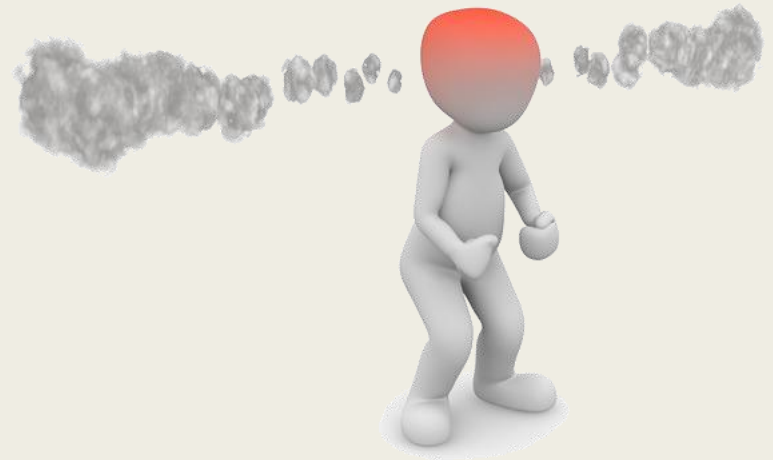
– Working Group on Korean Romanizer –

Hyoungbae Lee  
Korean Studies Librarian  
Princeton University

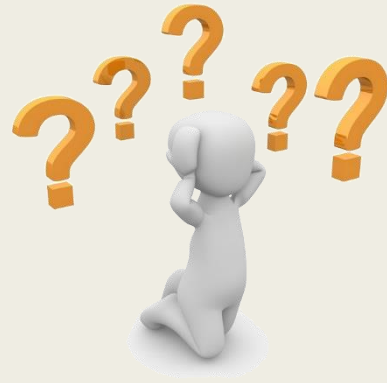


# Undeniable Headache

- Complicated sound changes
- Hybrid word division
- Tons of exceptions



# Déjà Vu



## Romanization

### ■ Rules

- 웃어 = us<sup>ö</sup>
- 식용 = sig<sup>y</sup>ong

### ■ Exceptions

- 웃어른 = ud<sup>ö</sup>r<sup>ü</sup>n
- 음식용 = ümsin<sup>g</sup>nyong

## Grammar

### ■ Rules

- *kid / kids*
- *watch / watched*

### ■ Exceptions

- *child / children*
- *see / saw*

# No Master Key, After All

- No romanization or pronunciation dictionaries
- No perfect natural language processing solutions (yet)
- No affordable super computers (yet)



# A Tool, Maybe



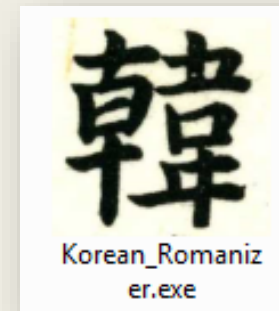
- Autolt v3 (from [www.autoitscript.com](http://www.autoitscript.com))
  - *Easy to learn BASIC-like syntax*
  - *Simulate keystrokes and mouse movements*
  - *Manipulate windows and processes*
  - *Interact with all standard windows controls*
  - *Scripts can be compiled into standalone executables*

# Autoit, Inside & Outside

## ■ Coding with AutoIt3

## ■ Executable Program

```
1384 Local $Rule[19][2] = [{"i10","ss"}, {"i11",""}, {"i12","c"}, {"i13","cc"}, {"i14",
1385 For $i = 0 To Ubound($Rule, 1) - 1
1386 $Result = StringRegExpReplace($Result, "\Q & $Rule[$i][0] & \E", $Rule[$i][
1387 Next
1388 Local $Rule[21][2] = [{"m20","i"}, {"m10","way"}, {"m11","oy"}, {"m12","yo"}, {"
1389 For $i = 0 To Ubound($Rule, 1) - 1
1390 $Result = StringRegExpReplace($Result, "\Q & $Rule[$i][0] & \E", $Rule[$i][
1391 Next
1392 Local $Rule[28][2] = [{"f20","ss"}, {"f21","ng"}, {"f22","c"}, {"f23","ch"}, {"f
1393 For $i = 0 To Ubound($Rule, 1) - 1
1394 $Result = StringRegExpReplace($Result, "\Q & $Rule[$i][0] & \E", $Rule[$i][
1395 Next
1396 $Rom = $Result
1397 $Rom = StringReplace($Rom, "#", "")
1398 $Rom = StringReplace($Rom, "~", "")
1399 $Rom = StringReplace($Rom, "i-75m-17f-26E", ".")
1400 $Rom = StringReplace($Rom, "i-75m-17", ",")
1401 $Rom = StringReplace($Rom, "f-14E", ":")
1402 $Rom = StringReplace($Rom, "i-75m-18f-11E", "!")
1403 $Rom = StringReplace($Rom, "f-9E", "?")
1404 $Rom = StringReplace($Rom, "E", "", -1, 1)
1405 If $NonKor > 0 Then
1406 $Rom = StringLeft($TargetKorOrig, $NonKor) & " " & $Rom
1407 EndIf
1408 ClipPut($Rom)
1409 EndFunc
```



# Showcase



- 김 영랑, 선우 휘, 강 빛나, 심 마르시아스
- 신여성들은 무엇을 꿈 꾸었는가?
- 눈물 없이 무지개는 피지 않는다 : 서울역 구두 닻이 출신 영어 교수의 살아 있는 자서전 (초판. 서울시 : 문화 행동, 1992)

# With Some Caveats

- Manual word division first
- Resolve ambiguities
- ALWAYS proofread





What Is Irreplaceable Is ...



# Because Computers Can't Tell

- Whether 삶은 계란 is
  - *salmŭn* kyeran “A boiled egg”
  - or
  - *sam ŭn* kyeran “Life is like an egg”
  
- Whether 문과 (文科) is
  - *mun kwa* “liberal arts”
  - or
  - *mun'gwa* “civil service examination”

# Working Group on Korean Romanizer

- Members
  - Hyoungbae Lee (Princeton, Chair)
  - Hyunjoo Eom (UBC)
  - Eunseung Oh (Stanford)
  - Jee-Young Park (Chicago)
  
- Tested and analyzed all the examples in ALA-LC Korean romanization rules.
  
- Strategic approach based on the analysis.
  
- Improvement and more features.

# Online Reporting & Discussion

Word	Romanizer Conversion	Correct Romanization	Report Date	Added to Exceptions	Romnizer Update	Note
역음	yökküm	yökküm	2015. 9. 21	N/A	2015. 9. 21	
실증	silch'üing	silchüing	2015. 9. 21	Yes	2015. 10. 13	
옛이야기	yesiyagi	yennyiyagi	2015. 9. 22	Yes	2015. 10. 13	
유권자	yugwönja	yukwönja	2015. 9. 28	Yes	2015. 10. 13	
기법	kiböp	kipöp	2015. 10. 5	Yes	2015. 10. 13	
대장간	taejanggan	taejangkan	2015. 10. 6	Yes	2015. 10. 13	
시골집	sigoljip	sigolchip	2015. 10. 6	Yes	2015. 10. 13	
단월대	Tanwöldae	Tanwöltae	2015. 10. 6	Yes	2015. 10. 13	
八達門	p'aldalmun	p'altalmun	2015. 10. 7	Yes	2015. 10. 13	
上八潭	sangp'aldam	sangp'altam	2015. 10. 7	Yes	2015. 10. 13	
강가 (near the river)	kangga	kangka	2015. 10. 7	Yes	2015. 10. 13	인도 관련 책에는 Kangga로 표기해야 하는 단어가 있으니 주의가 필요
팔당댐	p'aldangdaem	p'altangdaem	2015. 10. 7	Yes	2015. 10. 13	발음은 [팔쌍댐]이 맞으므로 '팔당 p'altang'을 예외로 등록 (LEE)
물가	mulga	mulka	2015. 10. 7	Yes	2015. 10. 13	price 혹은 waterside 에 따라 발음이 mulga / mulka 달라지는데 이런 경 >> 두 가지 다 발음이 같이 mulka입니다. 예외로 등록하였습니다. 그런데 같은 글자가 뜻에 따라 발음이 달라지면 정말로 방법이 없습니 (예) 알지(alchi) 모른다 vs. 알지(alji) 못한다
최고점	ch'oegojöm	ch'oegochöm	2015. 10. 8	Yes	2015. 10. 13	
딸감	ttaelgam	ttaelkam	2015. 10. 9	Yes	2015. 10. 13	
일등	ildüing	iltüing	2015. 10. 15	Yes	2015. 12. 1	일 <sup>ㅅ</sup> 등으로 입력하면 iltüing으로 나오는데 <sup>ㅅ</sup> 를 입력해서 경음화 현상을 나타낼수 있으면 report 할 필요 없는지요. 예를 들어 안다는 안 <sup>ㅅ</sup> 다로 입력하면 anta로 나오구요 뜻에 따라 발음이 달라지는 경우라 (know vs. hug) 리포트를 하지 않았습니. 일등은 예외로 등록해야 할것 같습니다 (oh) >> 일등은 ildüing으로 발음이 나는 경우가 없으므로 예외로 등록할 필요가 있습니다. '안다'의 경우는 말씀하신 대로 발음을 구별할 방법이 없으므로 리포트 해 주실 필요가 없구요 (Lee)
팔자	p'alja	p'alcha	2015. 10. 21	Yes	2015. 12. 1	팔자(Let's sell)이 있긴 하지만 레코드에 나올 가능성이 희박함(Lee)

# Try Demo Version

- <http://tinyurl.com/CEALDEMO>
  - Highlight Korean text and,
    - CTRL+ALT+K : for general use
    - CTRL+SHIFT+K : for personal names
  - CTRL+ALT+o/u : ö/ů
  - CTRL+SHIFT+\ : ‡
  - Works only on Windows PC
  - Will stop working after 30 days (Demo)
  - May require administrator permission  
(Right mouse click > Run as administrator)

