
Life's Good
SAMSUNG
Korean in 10 minutes
Jacob Lozano
SK hynix
안녕하세요
KIA
HYUNDAI

Language Profile

Southern Theory : Austronesian
- Fijian, Hawai'ian, Sunda, Tagalog …1248 (311M)

Northern Theory : Altaic
- Japanese, Turkish, Uzbek … 66 (250M)

National Language of North and South Korean
- Dialectal differences P'yöngyang & Seoul
  - China Japan, Russia, Thailand, Singapore …
  - ~78 Million world wide (62 in Korea)

Original Scripts
- Hyangchal (향찰/鄕札), Gugyeol (구결/口訣) and Idu (이두/吏讀).

Current Scripts
- Invented 1444, promulgated 1446
  - Low status, women, children and uneducated
- 19th & 20th century Hanja and Hangeul popular up
- 1945 Chinese Characters insignificant
- N. 1949-1960 hanjaless
- S. debate of role, used in official & academic papers

Hangul (한글) Chosŏn'gŭl (조선글)

The Korean Alphabet System Including Diphthongs And Double Consonants

CONSONANTS

ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅊ
g,k	n	d,t	r,l	m	b,p	s	ng	j	ch

ㅋ	ㅌ	ㅍ	ㅎ
k	t	p	h

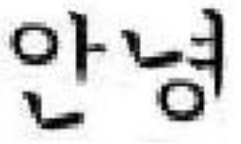
ㄲ	ㄸ	ㅃ	ㅆ	ㅉ
kk	tt	pp	ss	jj

VOWELS

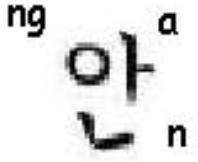
ㅏ	ㅑ	ㅓ	ㅕ	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	ㅣ
a	ya	eo	yeo	o	yo	u	yu	eu	i

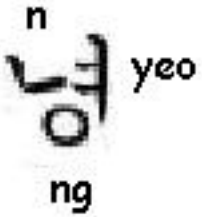
ㅐ	ㅒ	ㅔ	ㅖ	ㅘ	ㅙ	ㅚ	ㅝ	ㅞ	ㅟ	ㅢ
ae	yae	e	ye	wa	wae	oe	wo	we	wi	u

	g	n	d	r	m	b	s		j	ch	k	t	p	h
	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ae	gae	nae	dae	rae	mae	bae	sae	ae	jae	chae	kae	tae	pae	hae
ㅐ	개	내	대	래	매	배	새	애	재	채	캐	태	패	해
yae	gyae	nyae	dyae	ryae	myae	byae	syae	yae	jyae	chyae	kyae	tyae	pyae	hyae
ㅒ	걔	냬	댸	럐	먜	뱨	섀	얘	쟤	챼	컈	턔	퍠	햬
e	ge	ne	de	re	me	be	se	e	je	che	ke	te	pe	he
ㅔ	게	네	데	레	메	베	세	에	제	체	케	테	페	헤
ye	gye	nye	dye	rye	mye	bye	sye	ye	jye	chye	kye	tye	pye	hye
ㅖ	계	녜	뎨	례	몌	볘	셰	예	졔	쳬	켸	톄	폐	혜
wa	gwa	nwa	dwa	rwa	mwa	bwa	swa	wa	jwa	chwa	kwa	twa	pwa	hwa
ㅘ	과	놔	돠	롸	뫄	봐	솨	와	좌	촤	콰	톼	퐈	화
oe	goe	noe	doe	roe	moe	boe	soe	oe	joe	choe	koe	toe	poe	hoe
ㅚ	괴	뇌	되	뢰	묀	뵈	쇠	외	죄	최	쾨	퇴	푀	회
wae	gwae	nwae	dwae	rwae	mwae	bwae	swae	wae	jwae	chwae	kwae	twae	pwae	hwae
ㅙ	괘	놰	돼	뢔	뫠	봬	쇄	왜	좨	쵀	쾌	퇘	퐤	홰
wi	gwi	nwi	dwi	rwi	mwi	bwi	swi	wi	jwi	chwi	kwi	twi	pwi	hwi
ㅟ	귀	뉘	뒤	뤼	뮈	뷔	쉬	위	쥐	취	퀴	튀	퓌	휘
wo	gwo	nwo	dwo	rwo	mwo	bwo	swo	wo	jwo	chwo	kwo	two	pwo	hwo
ㅝ	궈	눠	둬	뤄	뭐	붜	숴	워	줘	춰	쿼	퉈	풔	훠
we	gwe	nwe	dwe	rwe	mwe	bwe	swe	we	jwe	chwe	kwe	twe	pwe	hwe
ㅞ	궤	눼	뒈	뤠	뭬	붸	쉐	웨	줴	췌	퀘	퉤	풰	훼
ui	gui	nui	dui	rui	mui	bui	sui	ui	jui	chui	kui	tui	pui	hui
ㅢ	긔	늬	듸	릐	믜	븨	싀	의	즤	칙	킈	틔	픠	희



"Annyeong"



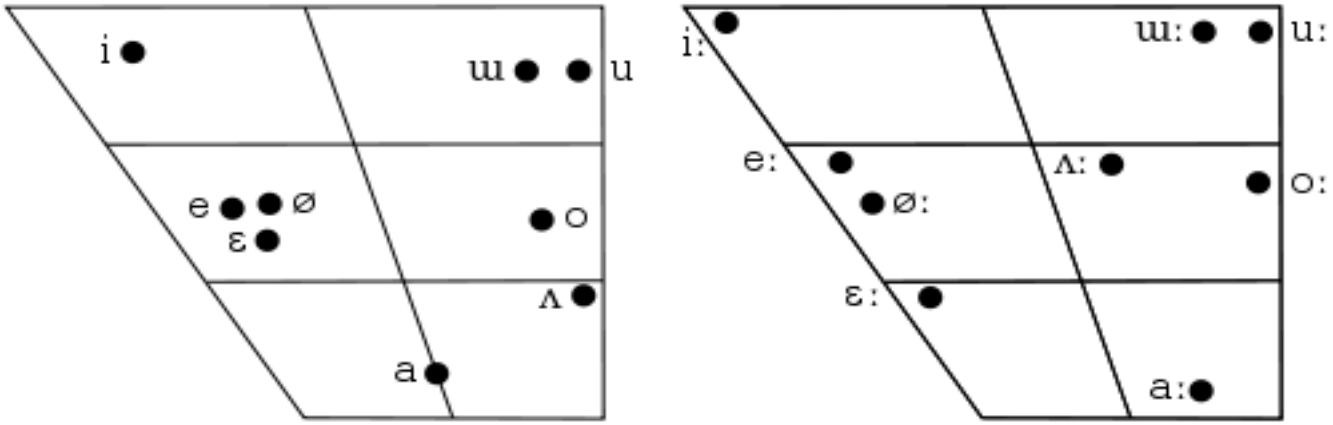


Phonology

- Voiceless, unaspirated, aspirated, glottalized
- 100+ obsolete cons/vowels

		Bilabial	Alveolar	Post-alveolar	Velar	Glottal
Nasal		ㅁ /m/	ㄴ /n/		ㅇ /ŋ/[19]	
Plosive and Affricate	plain	ㅂ /p/	ㄷ /t/	ㅈ /t͡ɕ/	ㄱ /k/	
	tense	ㅃ /p͈/	ㄸ /t͈/	ㅉ /t͈ɕ/	ㄲ /k͈/	
	aspirated	ㅍ /pʰ/	ㅌ /tʰ/	ㅊ /t͡ɕʰ/	ㅋ /kʰ/	
Fricative	plain		ㅅ /sʰ/			ㅎ /h/
	tense		ㅆ /s͈/			
Liquid		/w/[1]	ㄹ /l/	/j/[1]		

Hangul	ㅂ	ㄷ	ㅈ	ㄱ	ㅃ	ㄸ	ㅉ	ㄲ	ㅍ	ㅌ	ㅊ	ㅋ	ㅅ	ㅎ	ㅆ	ㅁ	ㄴ	ㅇ	ㄹ
RR	*b*	*d*	*j*	*g*	*pp*	*tt*	*jj*	*kk*	*p*	*t*	*ch*	*k*	*s*	*h*	*ss*	*m*	*n*	*ng*	*r, l*
IPA	p	t	t͡ɕ	k	p͈	t͈	t͡ɕ͈	k͈	pʰ	tʰ	t͡ɕʰ	kʰ	s	h	s͈	m	n	ŋ	ɾ, l



Hangul	ㅣ	ㅔ	ㅚ	ㅐ	ㅏ	ㅗ	ㅜ	ㅓ	ㅡ	ㅢ	ㅖ	ㅒ	ㅑ	ㅛ	ㅠ	ㅕ	ㅟ	ㅞ	ㅙ	ㅘ	ㅝ
RR	*i*	*e*	*oe*	*ae*	*a*	*o*	*u*	*eo*	*eu*	*ui*	*ye*	*yae*	*ya*	*yo*	*yu*	*yeo*	*wi*	*we*	*wae*	*wa*	*wo*
IPA	i	e	ø	ɛ	a	o	u	ʌ	ɯ	ɯi	je	jɛ	ja	jo	ju	jʌ	ɥi	we	wɛ	wa	wʌ

Morphology

Sentence : SOV

Nouns & Pronouns:

- no articles, gender/number marking
- Post-particles
  - Nominative, genitive, dative, accusative
  - Instrumental, locative, comitative
- Numeral Classifiers jang (장)
- I ‘this’ , ku ‘that’, ce ‘over there’

Verbs:

- No num/gen agree but polite
- Inflected base + ending (400+)
- Finite : honorific, tense, aspect, modal, formal, mood
- Tense
  - Marked: completed action
  - Unmarked: present
- Passive and Causative, +suffix
- Mood:
  - Declarative, interrogative, imperative, cohortative. Final of finite

a. 가 – 었 – 다

ka-ess-ta

GO-past-decl

b. 가 – 시 – 었 – 다

ka-si-ess-ta

GO-honorific-past-decl

c. 가 – 기 – 가

ka-ki-ka

GO-nominalizer-nom

d. 가 – 시 – 었 – 기 – 에 – 는

ka-si-ess-ki-ey-nun

GO-honorific-past-nominalizer-TO-topic

Declarative	**Ka-pni-ta**	**He is going**
interrogative	ka-pni-k’a	Is he going
imperative	Ka-la	Go.!
coharative	Ka-ca	Let’s go

MT Work : Morpheme tagger - spelling

Penn Tree Bank 54k

(3) 권한을 누가 갖고 있지?

kwenhan-ul-nwu-ka-kacko-iss-ci

AUTHORITY-acc WHO-nom HAVE BE-int

'Who has the authority?'

(4) (S (NP-OBJ-1 권한/NNC+을/PCA)
(S (NP-SBJ 누구/NPN+가/PCA)
(VP (VP (NP-OBJ *T*-1)
가지/VV+고/EAU)
있/VX+지/EFN))
?/SFN)

a. 동 ᄒᆡ/NN ⇒ 동 ᄒᆡ

tonghay *tonghay*

'east sea'

b. 동 ᄒᆡ/VV ⇒ 동 하 어

tonghay *tonghae*

'move'

Table II. Evaluation of spelling recovery

Process	Accuracy (%)
POS tagging for spelling recovery	92.03
Spelling recovery	96.33
Spelling recovery assuming 100% POS tagging	98.62

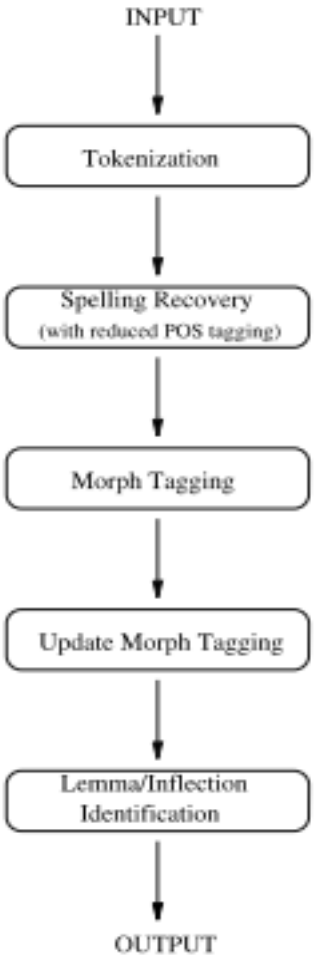


Figure 1. Overview of the tagger.

http://www.sfu.ca/~chunghye/papers/mt18-4-2.pdf

MT Work: Morpheme Tagger – 2/3 morph tags

165 complex tags 81%

Table III. Examples of complex tags

NNC
NNC+PCA
NNC+CO+EFN+PAD
NPR+PAD+PAU
NNU+PAU
VJ+EPF+EAN
VV+EPF+EFN
VV+EPF+EFN+PAU
VX+EPF+EFN+PAU

Few tag meaning

NNC common
CO copula
EFN type
PAD adveribial postposition
EPF tense

594 templates 93%

Table IV. Examples of inflectional templates

었습니다 *esssupnita*	VV+EPF+EFN
었으니까 *essunikka*	VV+EPF+ECS
었으며 *essumye*	VV+EPF+ECS
었으므로 *essumulo*	VV+EPF+ECS
었음에 *essumey*	VV+EPF+ENM+PAD
까지는 *kkacinun*	NNC+PAD+PAU
대로만 *tayloman*	NNC+PAD+PAU
로부터의 *lopwuteuy*	NNC+PAD+PCA

a. Input:
제가 관측 사항을 보고하였습니다.

Cey-ka kwanchuk sahang-ul pokoha-ess-supnita.

I-nom OBSERVATION ITEM-acc REPORT-past-decl

'I reported the observation items.'

b. Output:
제가/NPN+PCA 관측/NNC 사항을/NNC+PCA
보고하였습니다/VV+EPF+EFN ./SFN

a. Input:
제가/NPN+PCA 관측/NNC 사항을/NNC+PCA
잊었습니다/NNC

Cey-ka kwanchuk sahang-ul ic-ess-supnita.

I-nom OBSERVATION ITEM-acc FORGET-past-decl

'I forgot the observation items.'

b. Output:
제가/NPN+PCA 관측/NNC 사항을/NNC+PCA
잊었습니다/VV+EPF+EFN ./SFN

MT Work: Morpheme Tagger – 4 Lemma/inflection identification

Table V. Example entries from inflection and stem dictionaries

Inflection dictionary		Stem dictionary	
ㅂ시다 (*psita*)	EFN	간격 (*kankyek*)	NN
ㅂ시오 (*psio*)	EFN	간단하 (*kantanha*)	VJ
가 (*ka*)	PCA	간탁히 (*kanlyakhi*)	ADV
겠 (*keyss*)	EPF	간선 (*kansen*)	NNC
다 (*ta*)	EFN	간섭하 (*kansepha*)	VV
보다 (*pota*)	PAD	간섭 (*kansep*)	NNC
았 (*ass*)	EPF	간주하 (*kancwuha*)	VV
은 (*un*)	PAU	갈 (*kal*)	VV

(13) $abcd/p+x+y+z$

$\Downarrow$

$a/p+b/x+c/y+d/z$

Longest, until pos

(14) a. Input:

제가/NPN+PCA 관측/NNC 사항을/NNC+PCA
보고하였습니다/VV+EPF+EFN ./SFN

b. Output:

제/NPN+가/PCA 관측/NNC 사항/NNC+을/PCA
보고하/VV+었/EPF+습니다/EFN ./SFN

MT Work: Morpheme Tagger – Error and conclusion

Table VIII. Summary of error analysis

Type of error	Number	(%)	Gold	Test
Tagging	174	63	정직하/VJ+지/ECS *cengcikha-ci* HONEST-connective 'be honest'	정직하/VV+지/ECS *cengikha-ci*
Spelling recovery	57	21	피우/VV+ㄹ/EAN *piwu-l* EMINATE-adnominal 'which will eminate'	피울/NNC *piwul*
Ambiguity	16	6	부대/NNC+로/PAD *pwutay-lo* SQUAD-TO 'to the squad'	부/NNC+대로/PAD *pwu-taylo*
Missing NNC	15	5	학로/NNC *hakto* CADET 'cadet'	학/NNC+로/PAU *hak-to*
Missing template	11	4	훈련/NNC+이/CO+군/EFN *hwunlyun-i-kwun* DRILL-cop-decl 'be a drill'	훈련이군/NNC *hwunlyunikwun*
Missing inflection	4	1	소대급/NNC+에까지/PAD *sotaykup-eykkaci* PLATOON LEVEL-TOWARDS 'towards the platoon level'	소대급에/NNC+까지/PAD *sotaykupe-kkaci*

Table VI. Evaluation of the morphological tagger

Method	Precision (%)	Recall (%)
Treebank-trained	95.43	95.04
Treebank-trained^{+}	95.78	95.39

Table VII. Comparison with other morphological taggers

Reference	Accuracy (%)
Chan et al. (1998)	97.0
Yoon et al. (1999)	94.7
Lim et al. (1997)	94.8

MT Scores

Corpora

- Penn Tree Bank 34 files
- NIST 20 files
- SWRC KAIST corpus – korean-english-chinese 60k sent
- Speech ocean – korean-english-chinese-japanese 200k pairs

Some difficulties

- Borrowed words 뜨락또르 ttŭrakttorŭ Ru. трактор (traktor)
  트랙터 teuraekteo En. tractor tractor
- Difference in vocab 강냉이 kangnaeng-i 옥수수 oksusu corn

MT SCORE

English true

		BLEU-4 (mteval-v13a)			IBM BLEU (bleu-1.04)			NIST (mteval-v13a)			TER (tercom-0.7.25)			METEOR (meteor-0.7)		
SiteID	System	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB
CMU	CMU_kor2eng_primary_un	0.1113	0.1172	0.1049	0.1110	0.1171	0.1044	5.4745	5.5402	5.1113	0.7856	0.7824	0.7890	0.3974	0.4122	0.3810
KUNLPL	KUNLPL_kor2eng_primary_un	0.1118	0.1173	0.1052	0.1119	0.1174	0.1054	5.3482	5.1900	5.2084	0.9031	0.9126	0.8929	0.4397	0.4434	0.4356
SAIC	SAIC_kor2eng_primary_un	0.0943	0.0962	0.0923	0.0942	0.0962	0.0921	5.2579	5.1684	5.0722	0.8606	0.8629	0.8582	0.4252	0.4315	0.4183
UVA	UVA_kor2eng_primary_un	0.0679	0.0665	0.0693	0.0679	0.0665	0.0693	2.2916	2.3803	2.0755	0.7438	0.7520	0.7351	0.2915	0.3022	0.2796

Korean true

		BLEU-4 (mteval-v13a)			IBM BLEU (bleu-1.04)			NIST (mteval-v13a)			TER (tercom-0.7.25)			METEOR (meteor-0.7)		
SiteID	System	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB	Overall	NW	WB
CMU	CMU_kor2eng_primary_un	0.1047	0.1124	0.0963	0.1044	0.1124	0.0957	5.2021	5.3566	4.7619	0.7982	0.7881	0.8089	0.3815	0.3985	0.3626
KUNLPL	KUNLPL_kor2eng_primary_un	0.1014	0.1076	0.0943	0.1015	0.1077	0.0944	5.0712	5.0206	4.8328	0.9182	0.9208	0.9154	0.4189	0.4313	0.4052
SAIC	SAIC_kor2eng_primary_un	0.0881	0.0905	0.0854	0.0878	0.0902	0.0851	4.9513	4.9804	4.6584	0.8780	0.8720	0.8844	0.4049	0.4157	0.3929
UVA	UVA_kor2eng_primary_un	0.0608	0.0609	0.0605	0.0609	0.0610	0.0606	2.1462	2.2334	1.9402	0.7525	0.7602	0.7442	0.2775	0.2894	0.2643

Resources

Penn Korean Tree Bank
http://www.newdesign.aclweb.org/anthology/Y/Y02/Y02-1007.pdf

Morpheme tagger
http://www.sfu.ca/~chunghye/papers/mt18-4-2.pdf

KAIST Corpus
http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus

Speech Ocean
http://www.speechocean.com/en-Text-Corpora/696.html

Ethnologue
http://www.ethnologue.com/language/kor

Ominglot
http://www.omniglot.com/writing/korean.htm

About world languages
http://aboutworldlanguages.com/korean

감사합니다

gamsahabnida