

# **Annotation of Korean Learner's Corpora for Particle Error Detection**

## **Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs Calico Pre-Workshop 2008**

Seok Bae Jang (Georgetown University)  
Sun-Hee Lee (Wellesley College)  
Sang-kyu Seo (Yonsei Univeristy)

## **Outline**

1. Goals
2. Arguments
3. Background
  - 3.1. Properties of Korean
  - 3.2. Why Are Particles Important for a Learner Corpus?
4. Particle Annotation Project
  - 4.1. Data Collection Method: 10,000 Eojeol Learner Corpus
  - 4.2. Statistical Properties of the Learner Corpus (Heritage vs. Foreign)
  - 4.3. Error Type Classification and Annotation Scheme
  - 4.4. Example of Error Tagged Corpus
5. Findings and Relevant Issues

## Goals of Study

### ▲ Short-term Goal

1. Provide an annotation scheme that can be used as a gold standard for marking up particle errors in Korean learners' corpora.
2. Provide a useful resource for Korean language teaching, evaluation, second language acquisition, and heritage language learning.
3. Determine useful properties of a resourceful mark-up system that can facilitate automatic error detection as well as information extraction.

### ↑ Long-term Goal

Our long-term goal is to build up the automatic writing tutor system for Korean learners

## Arguments

1. Particle error tagging needs to handle multiple errors, which overlap in the same or partly same position.
2. Our error annotation study indicates that Korean Heritage learners show distinct properties from foreign learners in terms of particle errors.
  - Learner information needs to identify learners' language backgrounds so that it can be used for learner error analysis and other computational processes.
3. Reusability of Annotation  
Error mark-up can be used not only for identifying errors but also for providing more efficient feedback depending on different learner levels and background.

## Background

1. Korean belongs to the Agglutinative Language Type.
2. Verbs conjugate by adding (several) endings; nouns can separate from following particles, but morphologically they are tightly combined.
3. Morphosyntactic Combinations of Noun + Particle  
Similar to preposition + noun phrase except spacing is not allowed between a noun and a particle. ↑  
*학교-에서 내가 다니-는 학교-에서*  
*school-at I-Nom go-REL school-at*
4. Spacing is a big headache for Korean language processing and learning.
5. The Korean space-based unit, the *Eojeol*, corresponds to word units in English.

## Why Are Particle Errors Important in Korean?

1. Korean particles are hard for Korean Language Learners.  
Ko et al. (2004) – tagged corpus is/corpora are not available  
– Error analysis with 100,000 eojeol learner corpus.  
**lexical items (28.3%) > particles (24.4%) > misspelling (20.8%) > verbal endings(16%)**  
– Particle errors appear across different proficiency levels.
2. Particle errors are related to a learner's native language.
3. In writing, particle omission is restricted.

### Classification Particles

#### [1] Lexical Case

structural case: nominative(ka/i), accusative (ul/lul)

inherent case: dative, goal, locative, instrument, ... et.

#### [2] Discourse/Modal Particles: topic (un/nun), delimiters (to, man)

Particles can be stacked up. Combinations of [1] and [2]

## Building Tagged Learner Corpora

### ► Original Corpus

100 writing samples of Korean learners at Yonsei Korean Language Institute and Wellesley College

	Heritage	Non-Heritage	Total
Beginner	25	25	50
Intermediate	25	25	50

[1] Tagged Particle Errors were hand tagged.

[2] Learner background was carefully sorted, but consistent learner information was hard to maintain due to lack of information.

## An Importance of Header Information

- ◎ Learner information needs to be more carefully marked for more accurate data analysis and efficient feedback.
- ◎ In particular, student background in the target language provides different patterns of language learning in addition to proficiency levels. **Heritage vs. Non-Heritage**
- ◎ Genre [letter/diary/essay/formal writing/...] and mode [spoken/written] information plays a crucial role in determining speech levels in Korean.
- ◎ In addition, learner questionnaires need to be kept as supporting information.

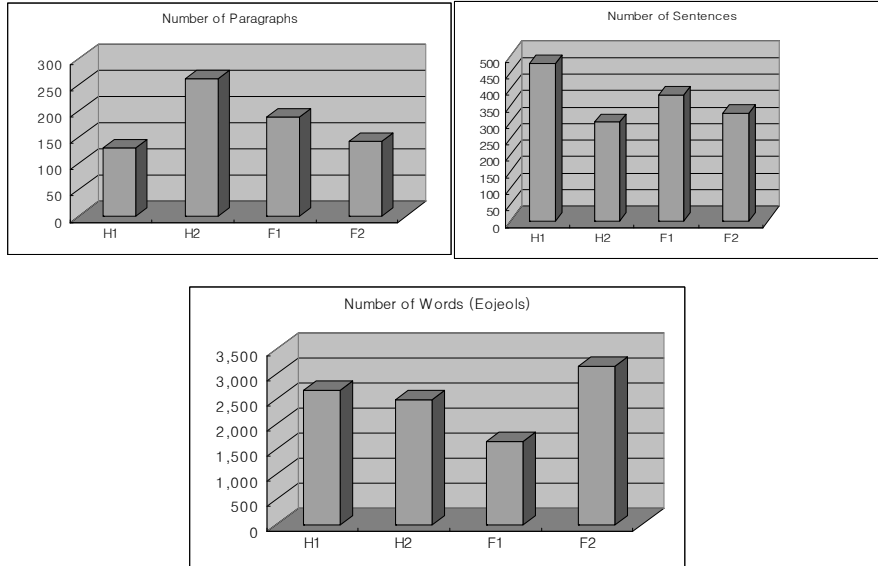
## Header Mark-up

[1]	<set id =“F2-001”>	Learner ID
[2]	<place> Korea </place>	Place of Data Collection
[3]	<date> 2004.4 </date>	Date
[4]	<name>***** ****</name>	Name
[5]	<nationality>USA</nationality>	Nationality
[6]	<level> 4 </level>	Proficiency level
[7]	<korean speaking parent>none </korean speaking parent>	Heritage info
[8]	<period of exposure> 0 </period of exposure>	Degree of exposure to Korean language
[9]	<school> KLI </school>	School
[10]	<period> 2 years </period>	Year of Study
[11]	<foreign language> French</foreign language>	Other foreign language
[12]	<gender>M </gender>	Gender
[13]	<age> 40 </age>	Age
[14]	<genre> essay </genre>	Genre
[15]	<style> informal polite </style>	Style
[16]	<topic> 한국 생활</topic>	Topic

## Statistical Properties of the Learner Corpus

<b>&lt;Heritage Beginner&gt;</b>		<b>&lt;Foreigner Beginner&gt;</b>	
Number of Sets:	25	Number of Sets:	25
Number of Paragraphs:	129	Number of Paragraphs:	188
Number of Words:	2,669	Number of Words:	1,659
Number of Sents:	476	Number of Sents:	381
<b>&lt;Heritage Intermediate&gt;</b>		<b>&lt;Foreigner Intermediate&gt;</b>	
Number of Sets:	25	Number of Sets:	25
Number of Paragraphs:	261	Number of Paragraphs:	142
Number of Words:	2,496	Number of Words:	3,163
Number of Sents:	299	Number of Sents:	327

## Heritage Learners vs. Foreign Learners



## Heritage Learners vs. Foreign Learners

Heritage VS Foreign	<Heritage>	Percent	<Foreign>	Percent
<b>Number of Sets</b>	<b>50</b>		<b>50</b>	
<b>Number of Paragraphs</b>	<b>390</b>	<b>54.17</b>	<b>330</b>	<b>45.83</b>
<b>Number of Sentence</b>	<b>775</b>	<b>52.26</b>	<b>708</b>	<b>47.74</b>
<b>Number of Words</b>	<b>5,165</b>	<b>51.72</b>	<b>4,822</b>	<b>48.28</b>
<b>Words per Sentence</b>	<b>6.66</b>		<b>6.81</b>	

## Beginners vs. Intermediates

	<Beginner>	Percent	<Intermediate>	Percent
<b>Number of Sets</b>	<b>50</b>		<b>50</b>	
<b>Number of Paragraphs</b>	<b>317</b>	<b>44.03</b>	<b>403</b>	<b>55.97</b>
<b>Number of Sentence</b>	<b>857</b>	<b>57.79</b>	<b>606</b>	<b>42.21</b>
<b>Number of Words</b>	<b>4,328</b>	<b>43.34</b>	<b>5,659</b>	<b>56.66</b>
<b>Words per Sentence</b>		<b>5.05</b>		<b>9.04</b>

## Error Type Classification and Annotation Scheme

### ►6 Types Particle Errors ◀

Type 1: Omission

Type 2: Replacement

Type 3: Addition

Type 4: Malformation

Type 5: Paraphrasing

Type 6: Spacing

1. Given particle error types include various levels of errors: morphological, lexical, syntax, discourse
2. Multiple error tagging was allowed.
3. Error tagging was matched with feedback (reconstruction)
4. Formality was considered for particle error mark-ups.

## Type 1: Omission

우리-Ø 편한            미국 생활-Ø            떠나서            여기-Ø 왔어요  
uri    phyeonha-n    mikuk saenghwal    tteonaseo    yeoki wasseoyo  
we    comfortable    American life            left            here    came  
'We left our comfortable life in America and came here.'

N.B. Case Dropping is more restricted in written Korean.  
Case Marker Omissions for intermediate foreigners:

Subject Marker: <b>i/ka</b>	total: 35
Object Marker: <b>ul/lul</b>	total: 36
Topic Marker: <b>un/nun</b>	total: 20

etc.

## Type 2: Replacement

한국-에서    온 후-에    한국 친구-를    많이 사귀고...  
hankuk-**eyse**    on hu-ey    hankuk chinku-lul    manhi sakwiko...  
Korea-from    come after Korean friends-Acc many  
'After I came to Korea, I met many Korean friends...'  
: **에서** eyse → **에** ey

선생님-한테            바른 말-으로            배우는 것-은            말해요,  
seonsaengnim-**hanthey**    paleun mal-ulo    payunun    keos-**un** malhayyo  
'It means learning right words from a teacher.'  
: **한테** hanthey → **께** kkey (honorific)

1. Particles are based on the subcategorization frame of verbs or adjectives.
2. Topic markers and other delimiters are based on semantic and discourse properties.
3. Honorific marking in Korean is associated with particles and verbal endings.

### Type 3: Addition

봄-에-는      꽃-이      피-는      걸      보니까      내-가...  
pom-ey-nun      kkoch-i      phi-nun      keo-l      po-  
nikka    nay-ka  
spring-in-top      flower-Nom bloom-Rel    thing-Acc    see-since I-  
Nom...  
'Since (I) see flowers bloom in spring, I...'

► Topic marker *nun* needs to be deleted.

1. Redundant particles need to be marked for deleting.
2. There are specific cases in which no particles are required.  
For this, annotation guidelines need to specify examples and conditions.  
: Lee & Park (2007)

### Type 4: Malformation

가족-이나    친구-이나    이웃-이      있어-서    행복할      수도  
있고.  
Kacok-ina    chinkwu-ina    iwus-i      issesseo    hangbokhal    swuto  
isso....  
Family-or    friend-or      neighbor-Nom    exist-since    happy      way  
exist  
'Since we have family, friends, or neighbors, we can be happy...'  
: **이나 ina → 나 na**

1. Many particles in Korean have phonological variation. The forms change depending on if the previous syllable ends with a consonant or a vowel.
2. Typos and wrong word forms belong to this type.
3. Spell checkers are expected to reduce the learner errors but still remaining errors will trigger more counts in this error type.

## Type 5: Paraphrasing

**행복-이라고      사람-마다      다르다**  
 Hangbok-ilako      salam-mata taluda.  
 happiness-Copular person-each different  
 'As for happiness, every person has different ideas.'  
**이라고 ilako → 은 un**  
 copular → topic marker

- [1] A particle can replace a morpheme or a phrase.  
→ Only this case was marked as paraphrasing.
- [2] Also a particle or the unit which a particle belongs to can be replaced by some phrase or a verbal ending.

**상-을      받-는      것-이 (→ 받게)      되면      기분-이      좋지**  
 sang-ul      pat-nun      kes-i (→ pat-key)      toymyon      kibun-i      cohci.  
 award-Acc receive-Rel thing-Nom (receive-E)      become      feeling-Nom good  
 'If I become to receive an award, I feel good.'

## Type 6: Spacing

- ▶ Forward spacing:  
Particles attach to the preceding noun without any space.

**큰 도시 에 서 는      때때로      길-이      혼잡해요,**  
 kun tosi ey se nun      ttayttaylo      kil-l      honjaphayyo.  
 big city at      Top      sometimes      road-Nom      crowded  
 'In a big city, the road is very crowded.'

- ▶ Backward spacing:  
Particles are separated from the following word by using space.

**예-를 들-어서      한국어-를      배우기-를      빼고는**  
 ye-lultul-ese      hankuke-lul      paywuki-lul      ppayko-nun  
 example-Acc take-if      Korean-Acc      learning-Acc      exclude-Top...'  
 'For example, excluding learning Korean....'

## Particle Error Tagged Learner Corpus

### Header Information

```
<set id="H1-002">
<head>
<date>2001/07/02</date>
<place>한국</place>
<name>***</name>
<nationality>KUSA</nationality>
<mother_lang>SENG</mother_lang>
<level>BE-L1</level>
<school>Pub-YON</school>
<period>14M</period>
<task_type>T4</task_type>
<gender>F</gender>
<age>20</age>
<occupation>STU</occupation>
<topic>4</topic>
</head>
```

## Particle Error Tagged Learner Corpus

```
<text>
<p>
<s><err errid="1">크리스마스(christmas)</err> 너무 조와합니다.</s>
<s>굉장히 빠른 시간이지만, 아직도 재밌습니다.</s>
<s><err errid="2">선물존는걸</err> 그리고 바다는걸 조와합니다.</s>
<s>친구하고 <err errid="3">선물가왔습니다.</err></s>
<s>크리스마스 <err errid="4">캐를가</err> <err errid="5">찬손</err> 드리고, <err errid="6">
  쿠키</err> 많이 만드셨습니다.</s>
</p>
<p>
<s><err errid="7">어머니</err> 맞시는 <err errid="8">음식</err> 만드셨습니다.</s>
<s><err errid="9">미국음식</err> <err errid="10">한국음식</err> 만드셨습니다.</s>
<s>목 <err errid="11">하고</err> 두부를 먹고, 미국 닭고기를 먹습니다.</s>
<s>한 <err errid="12">점시에서</err> 미국 <err errid="13">음식</err> 있습니다.</s>
<s>한 점시 <err errid="14">에서</err> <err errid="15">한국음식</err> 있습니다.</s>
</p>
<p>
<s>나 <err errid="16">한태</err> <err errid="17">딸개를</err> <err errid="18">맞시</err> 업
  습니다.</s>
</p>
</text>
```

## Error mark-up

```

<errlink relatedToError="6" type="omission" order="1" feedback="insert 를"/>
<errlink relatedToError="7" type="omission" order="1" feedback="insert 께서"/>
<errlink relatedToError="7" type="comment" feedback="Change the verb form 만듭습니
다:만드셨습니다"/>
<errlink relatedToError="8" type="omission" order="1" feedback="insert 을"/>
<errlink relatedToError="9" type="malformation" order="1" feedback="귀:과"/>
<errlink relatedToError="10" type="omission" order="1" feedback="insert 을"/>
<errlink relatedToError="12" type="omission" order="1" feedback="insert 는"/>
<errlink relatedToError="12" type="replacement" order="2" feedback="에서:에"/>
<errlink relatedToError="12" type="malformation" order="3" feedback="에서:에서"/>
<errlink relatedToError="13" type="omission" order="1" feedback="insert 이"/>
<errlink relatedToError="14" type="omission" order="1" feedback="insert 는"/>
<errlink relatedToError="14" type="replacement" order="2" feedback="에서:에"/>
<errlink relatedToError="14" type="malformation" order="3" feedback="에서:에서"/>
<errlink relatedToError="15" type="omission" order="1" feedback="insert 이"/>
<errlink relatedToError="16" type="replacement" order="1" feedback="한태:는"/>
<errlink relatedToError="16" type="malformation" order="2" feedback="한태:한테"/>
<errlink relatedToError="16" type="space_front" order="3" feedback="No space between
the preceding noun and the verb"/>

```

## Error Patterns

1. Overlapping Error Types    Example

2. Particle Errors and the Following Verbs. Example

3. Pseudo-Particles Composed of a Particle and a Predicate

-을 가지고	-에 의해서/의하면	-에 대해서/대한	-과 함께
-ul kajiko	-ey uyhaeseo	-ey tayhayse/tayhan	-kwa hamkkey
‘in terms of’	‘by’	‘about’	‘with’

4. Particle Errors in Unintelligible Phrases    Example

<s>나 한태 딸개를 맛시 업습니다. </s>  
 na hanthay ttalgaelul masssi epsupnita  
 I to strawberry-Acc tast-i not exist  
 'To me, strawberries are not tasty.'

- 1) 한태 *hanthay* → spelling error ----> 한테 *hanthey*
- 2) 한테 *hanthey* needs to be attached to the preceding noun.
- 3) 한테 *hanthey* needs to be replaced by the topic marker 는 *nun*

⊙ Rule ordering may need to be considered for feedback.

<s>...부모님이 나를 한국에 보냈어요.</s>  
 pwumonim-i na-lul hankuk-ey ponaysseyo.  
 parents[+Hon]-Nom I-Acc Korea-to sent.  
 'My parents sent me to Korea.'

- 1.0이 *i* needs to be changed into the honorific nominative 께서 *kkeyse*.
- 2.보냈어요 *po-nayss-eyo* needs to be 보내셨어요 *po-naysyess-eyo*.
3. 나 *na* needs to be changed into the humble form 저 *ce*

### Unintelligible input ?????

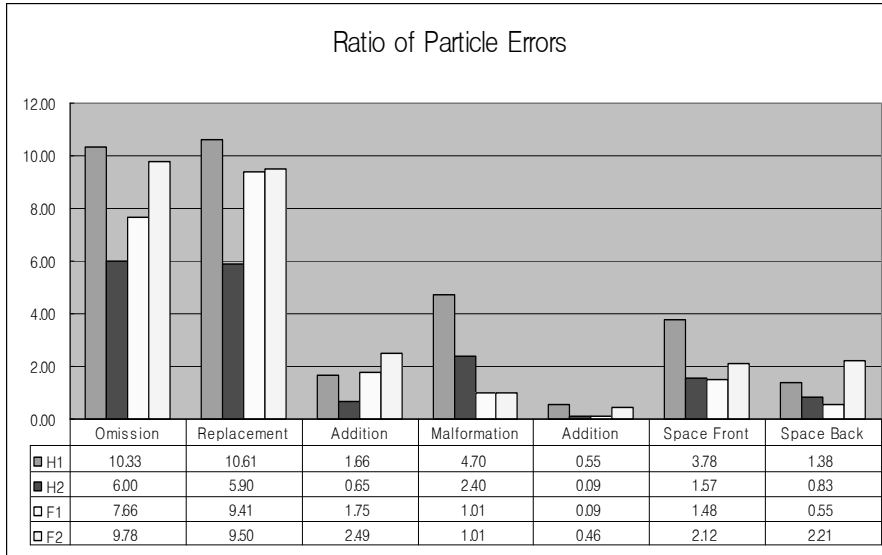
<s>한국의 문화를 내 인생-에세 살리고 싶습니다</S>  
 hankuk-uy munhwa-lul nay insayng-eysey salliko sipsupnita.  
 Korea-Gen culture-Acc my life -in alive want  
 'I want to make Korean culture alive in my life.'

<s>이모Ø 많이 도와주고 어머니 하고 아버지 사랑해요.</s>  
 imo Ø manhi towacuko emeni hako abeci salanghaeyo  
 ant a lot help mother and father love  
 'I am helping my aunt a lot and I love you, mother and father!'  
 'My aunt helps me a lot and I have you, mother and father!'

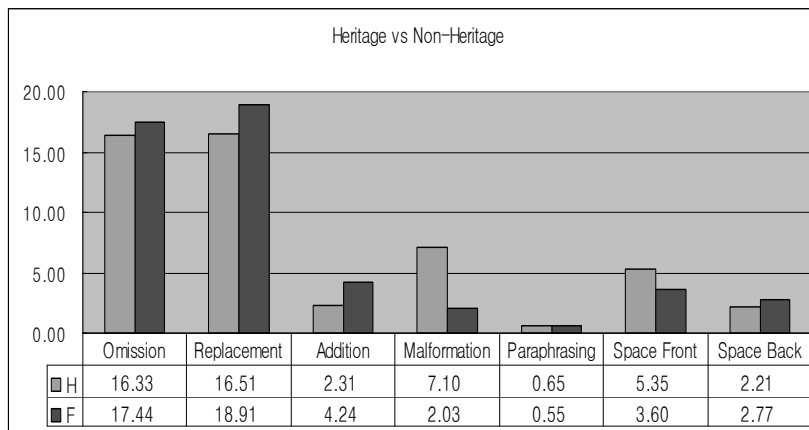
### Beginners: Heritage vs. Non-Heritage

	Heritage Beginner		Non-Heritage Beginner	
Omission	112	31.28	83	34.87
Replacement	115	32.12	102	42.86
Addition	18	5.03	19	7.98
Malformation	51	14.25	11	4.62
Paraphrasing	6	1.68	1	0.42
Space Front	41	11.45	16	6.72
Space Total	15	4.19	6	2.52
sum	358	100.00	238	100.00

## Particle Error Distributions



## Heritage vs. Non-Heritage



## Our Findings and Relevant Issues

### I. Error Difference between Heritage vs. Non-Heritage

[1] Difference in particle error types of heritage learners and non-heritage learners suggests that different approaches need to be taken for teaching and also for evaluation process.

#### [2] Heritage Language Learning

Second Language Acquisition + First Language Acquisition  
+ Teaching Methods + Evaluation Process

## Our Findings and Relevant Issues

### II. Error Tagging

1. Error tagging needs to be matched with the feedback process.
  - ✓ In particular, particle errors are closely associated with other components in a sentence. This needs to be considered in the feedback process.
2. Heritage learners show unexpected errors across different levels.
  - ✓ For heritage learners, user-oriented feedback will be more efficient than level-based feedback.
  - ✓ Well-defined error annotation guidelines are necessary for verbal ending and other types of errors on the based of user-oriented feedback.
3. Complex errors including space errors, lexical selection errors, etc. need to be ultimately incorporated in annotation system.
  - ✓ Rule ordering will be needed for providing correct feedback.

## Future Work

1. Include more specific details of annotation guidelines.
2. Train annotators and check inter-annotator agreement rate.
3. Expand error tagging with larger size learner corpus.
4. Compare the error patterns of Korean American learners with other heritage learner corpora
5. Develop a computational error tagging tool.
6. Elaborate error type analysis in connection with learner differences and feedback.
7. Investigate the empirical usage of tagged error corpus while pursuing to handle open end user inputs  
Challenge: serious lack of computational tools

감사합니다!

Thank You !