Annotation of Korean Learner Corpora for Particle Error Detection

Sun-Hee Lee
Wellesley College
Seok Bae Jang
Brigham Young University
Sang-Kyu Seo
Yonsei University

ABSTRACT
In this study, we focus on particle errors and discuss an annotation scheme for Korean learner corpora that can be used to extract heuristic patterns of particle errors efficiently. We investigate different properties of particle errors so that they can be later used to identify learner errors automatically, and we provide resourceful annotation guidelines. We present issues that are relevant to learner error annotation including how to classify particle error types, present correct tokens, mark overlapping error types, and so on. Accurate annotation of particle errors will be useful for extracting relevant error rules and will provide substantial benefits to the development of ICALL systems. We argue that it is necessary to link annotation with feedback procedures to increase the efficiency of systems. Furthermore, we investigate whether heritage learners and nonheritage learners generate different error patterns and discuss significant implications for heritage versus nonheritage language learning.

KEYWORDS
Learner Corpus, Particles, Annotation, Error, Korean

INTRODUCTION
Over the last decade, large corpora and computational tools, including morphological analyzers, part-of-speech taggers, syntactic parsers, and so on, have been developed in Korea. The 21st Century Sejong Project, a Korean national corpus project, has especially made it possible to obtain free access to various corpora and computational tools for language research. In line with this, Korean learner corpora have begun to be developed by researchers who have recognized their importance in language learning and teaching. Annotated learner corpora are known to provide valuable information about language learners in teaching environments, language acquisition, and language variation. They are also useful for developing automatic tutor systems in which computers instead of humans detect errors and provide feedback to language learners. Specifically, error mark-up can contribute to the selection of features that are useful for identifying errors and testing the accuracy of diagnostic processes in tutor systems for writing.

Despite the value of learner corpora, their usability remains a challenge in relevant fields due to a lack of resourceful mark-up guidelines and available computational tools. In this study, we begin to address this challenge by providing an annotated learner corpus for...
particle errors. This corpus can be used for developing an automatic error detection process, as well as for information extraction, because it provides subclasses of particle errors and error patterns. Our long term goal is to build an intelligent computer-assisted language learning (ICALL) system for Korean that provides automatic particle error detection and interactive feedback. As a stepping stone, we provide resourceful annotation guidelines for particle errors and investigate properties of the Korean learner corpora that may be useful for diagnosing learner errors. Furthermore, our analysis of particle error mark-ups suggests that heritage learners and nonheritage learners show different error patterns. The implication of such a finding is that distinct approaches may be more effective for each group of learners than a one-size-fits-all approach. Ultimately, we argue that it is essential to know what kind of error mark-ups are most useful for the feedback process; that is, there must be a link between error diagnosis and feedback procedures in order to enhance the efficiency of an automatic tutor system.

**PROPERTIES OF KOREAN AND PARTICLE ERRORS**

Korean is an agglutinative language. Verbs are conjugated by adding different verbal endings, and nouns can be morphosyntactically combined with different particles. The combination of a noun and a particle in Korean is similar to the combination of a preposition and a noun in English but follows inverse word order and has no spacing between the noun and the particle. For example, the combination of 학교 hakkyo ‘school’ and a locative particle 에서 eyse ‘at’ appears as 학교에서 hakkyoeyse ‘at school.’ In addition, a crucial difference between Korean and English is that subject and object functions appear with particles in Korean, whereas subject and object are distinguished by word order in English. Particles mark the grammatical function of nominals, such as subject and object, and add specific meanings including location, instrument, goal, and so forth. They can also provide discourse functions. In sum, particles in Korean can be divided into two different categories according to their functions. These categories and subcategories are shown in (1).

(1) **Classification of Particles**

---

<table>
<thead>
<tr>
<th>Lexical Case →</th>
<th>Structural Case:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominative (가/이 ka/i, -께서 kkeyse),</td>
</tr>
<tr>
<td></td>
<td>Accusative (을/를 ul/lul),</td>
</tr>
<tr>
<td></td>
<td>Genitive (의 uy)</td>
</tr>
<tr>
<td></td>
<td>Inherent Case:</td>
</tr>
<tr>
<td></td>
<td>Dative (에게 eykey),</td>
</tr>
<tr>
<td></td>
<td>Goal (으로/로 ulo/lo), 에 ey),</td>
</tr>
<tr>
<td></td>
<td>Locative (에서 eyse),</td>
</tr>
<tr>
<td></td>
<td>Instrument (으로/로 ulo/lo), etc. etc.</td>
</tr>
<tr>
<td>Discourse/Modal Particles →</td>
<td>Topic (은/는 un/nun), Delimiters (도 to, 만 man)</td>
</tr>
</tbody>
</table>

The first category includes lexical case, which can be subdivided into structural case and inherent case. Structural case marks the syntactic function of an argument (e.g., nominative, accusative, and genitive). Inherent case marks the semantic relations between an argument and a predicate (e.g., locative, instrument, and goal). The second category includes topic markers or delimiters whose properties are closely linked to discourse properties. Korean particles can be stacked one after the other, although there are morphosyntactic restrictions on certain combinations. For example, 너만을 ne-man-ul ‘you-only-accusative’ is composed of a pronoun 너 'you,' the delimiter 만 man ‘only,’ and the accusative particle 을 ul (for more details on Korean particles, see Lee, 2004; Nam & Ko, 1993).
Realization of lexical case is also closely related to predicates. For example, the honorific nominative particle *kkyese* marks subjecthood, whereas the main predicate takes the honorific morpheme *si* or a separate honorific form. In the following example, the predicate takes the honorific morpheme *시* *si* and the subject takes the honorific nominative case *kkeyse*.

(2) 어머니-께시 학교-에 가-시-다.
Mother-Nom school-to go-Hon-End

'Mother goes to school.'

Particle errors are one of the most frequent errors that Korean language learners produce. According to Ko et al. (2004), particle errors accounted for 28% of all errors made by beginning-level learners, that is, there were 3295 *eojeol* errors in a 100,000 *eojeol* learner corpus (*eojeol* refers to a unit set off by spaces, which corresponds to a word unit in English). In addition, the authors reported that particle errors were the second most frequent error type across six different levels of Korean learners. It is worth noting that in their study the major type of particle error varied for learners at different levels of proficiency even though there was a general decrease in errors as learners moved to higher levels. Low-level learners showed basic mistakes in selecting proper particle forms without awareness of the semantic connectivity between the correct form and erroneous forms. In contrast, (advanced) intermediate-level learners induced more errors by switching particles based on close semantic connectivity. The latter type of error is hard for a parser to capture since it involves subtle semantic and pragmatic properties. However, in order to develop a full-fledged error diagnosis, it is essential to incorporate these sophisticated linguistic features into particle error annotation. The details will be discussed below.

BUILDING A KOREAN LEARNER CORPUS WITH PARTICLE ERROR TAGGING

In this section, we present the general procedure for building a Korean learner corpus as well as relevant issues that we encountered. We also provide a classification of particle errors and useful linguistic information that can facilitate automatic error detection processes.

Data Collection Process and Importance of Learner Information

In order to prepare the corpus, we collected writing samples from 100 Korean language learners at Yonsei Korean Language Institute in Seoul, Korea, and Wellesley College in the US. By combining written assignments from both institutions, we achieved a corpus containing 50 texts from beginners at levels 1 and 2 and 50 texts from intermediate learners at levels 4 and 5. The beginner and intermediate learners’ data was further balanced into writing samples from 25 Korean heritage learners and 25 nonheritage learners. Texts were classified as belonging to heritage or nonheritage students based on learner background information provided by Yonsei Korean Language Institute and also obtained from the Wellesley College participants. The term ‘Korean heritage learner’ refers to Korean language learners who have grown up in Korean-speaking environments, although they do not necessarily speak the language. In general, their parents use Korean at home and communicate with their children in Korean. Heritage learners are capable of understanding Korean, but they may reply either in Korean or in English. Heritage learners in Korean classrooms do not know how to write Korean at all or have low-level writing skills.
The decision to compare heritage and nonheritage learners stems from the assumption that language learning is different for these two groups, as has been argued by Polinsky and Kagan (2007), Montrul (2004), among others. Indeed, our examination of the corpus provided evidence that particle errors in Korean are distinct for heritage learners and nonheritage learners.

We typed each of the handwritten samples into a text file format that preserved the original forms, including the errors, and stored learner background information in a header file attached to the raw text. Later, we added linguistic information on each particle error following our annotation schema.

In order to increase the usability of corpora, as well as the accuracy of data analysis and the efficiency of feedback systems that may ultimately operate according to learner level and background, it is crucial to maintain concrete, consistent information on learners. In particular, student background in the target language can be used to explicate differences in the language-learning process and error production. Genre information (e.g., whether the text is a letter, diary, essay, formal writing, etc.) and mode of information (e.g., whether the text reflects spoken or written information) both play a crucial role in determining speech levels in Korean. By adding the learner’s status as a heritage or nonheritage student and by assigning an indicator of the genre of the text into the header part of the text, we modified the existing header information that was used in Ko et al. (2004). Figure 1 shows an example of the header information attached to each of our samples. The bold-faced information represents information that distinguishes heritage learners and nonheritage learners and records the degree of exposure to the language as well as the genre specification. Collecting this information was essential for conducting applied research with the learner corpus.

Figure 1
Sample Header Information

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

Following Ko et al. (2004), we classified particle errors into six types: omission, addition, replacement, malformation, paraphrasing, and spacing. We examined the kinds of properties that exist across particle errors and considered how they should be marked to facilitate automatic error diagnosis.

Consider for a moment Japanese particles, which are similar to Korean particles. Nagata (1996) developed a system that corrects a learner’s Japanese input in English-Japanese translation exercises. Her system detects errors in particles by assuming a one-to-one correspondence between Japanese particles and English prepositions. However, her assumption only works within a narrowly restricted domain: parallel sentence representations in Japanese and English. If we consider the language specific properties of Korean, this approach seems to be rather dangerous because making this assumption will lead us to develop a system based on unnaturally translated Korean sentences. This is an undesirable side effect for language learning. In addition, it does not capture complex properties of particle errors beyond basic level errors that involve form mismatching. Appropriate annotation of particle errors will contribute to the development of more sophisticated systems that can process particle errors and provide automatic correction. Moreover, we argue that the annotation scheme of particle errors needs to be designed to mark up errors for both advanced-level learners as well as low-level learners and that different feedback needs to be offered according to learner levels. The importance of providing different intelligent feedback depending on learner levels has been increasingly emphasized in the communities of ICALL in particular and CALL in general. The properties of each error type are specified below.

[Type I] Omission

Particles are known to disappear in languages like Korean and Japanese, especially in speech. However, case dropping is more restricted in written Korean and in formal spoken Korean domains such as public debate, broadcast shows, lectures, and so forth. Even in speech, particle dropping rate is unexpectedly low and varies depending on properties of register (see Lee & Park, 2008). For Korean learners, particle omission is one of the most frequent errors across the different proficiency levels. Dropping too many particles sounds unacceptable, unnatural, or in many cases like baby talk. An example of particle omission is shown (3)

(3) 우리-Ø(←가) 편한 미국생활- Ø(←음) 떠나서 여기 왔어요.
   uri- Ø(←ka) phyeonha-n mikuksaenghwal(←ul) tteonaseo yeoki wasseoyo.
   ‘We left our comfortable life in America and came here.’

Here, the subject particle and the object particle have been omitted and the sentence is not natural. In order to enhance the proficiency of the target language and achieve successful language learning, it is crucial to capture omission errors and correct them. Because our learner corpus is composed of writing samples and our long-term project aims to develop an automatic writing tutor, we counted the maximum number of omission errors by assuming strict adherence to required particle use.
[Type II] Addition

(4) 봄에는 (delete) 꽃이 꽃은 걸 보니까 내가 ...
pom-ey-nun (delete) 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…'

In (4), the topic marker nun is redundant and should be marked as an addition. Furthermore, there are specific cases in which no particles are required, and yet learners may incorrectly add them. For example, in the fossilized idiomatic expression 그 사람-Ø/-'이 있잖아 ku salam-Ø/-'i isscanha 'as for that person,' the subject particle must be dropped.

In terms of Type I and II errors, the annotation guidelines need to specify empirical examples and conditions that describe necessary particle omissions and deletions so that they are not marked as errors. Otherwise, interannotator agreement will be particularly low for these two error types.

[Type III] Replacement

The third error type is particle replacement as in (5).

(5) 한국에서 (→에) 온 후에 한국 친구를 많이 사귀고 ...
hankuk-eyse(→ey) o-n hu-ey hankuk chinku-lul manhi sakwiko ...
Korea-from come-REL after Korean friends-Acc many met

'After I came to Korea, I met many Korean friends…'

The locative particle 에서 eyse cannot combine with the following verb form, 온 o-n. In Korean, a particle is attached to an argument based on the subcategorization frame of a verb or an adjective. Particles of inherent case can be combined with an adjunct and add specific lexical meaning to it. In contrast, topic markers and other delimiters are based on semantic and discourse properties. Error mark-ups for particle replacement need to consider the subcategorization frame of the following verb, lexical meaning, and discourse properties of other particles. In addition, we need to note that the subject/dative honorific marking in Korean is morphosyntactically associated with the preceding nominal and the following verb, as well as with sociolinguistic background.

[Type IV] Malformation

Malformation refers to errors in particle forms (or spelling). Some particles have phonological variants. For example, the form of a particle changes depending on whether the previous syllable ends with a consonant or a vowel. In (6), the phonological variant 이나 ina has been wrongly used instead of 나 na after a syllable ending with a vowel.

(6) 가족이나 친구이나 (→나) 이웃-이 있어서 행복할 수도 있고
kacok-ina chinkwu-ina(→na) iwus-I issesseo hangbokhal swu-to iss-ko ...
family-or friend-or neighbor-Nom exist-since happy way-also exist

'Since we have family, friends, or neighbors, we can be happy…'

For Type IV errors, spell checkers are expected to help reduce the number of the learner errors. However, our data was originally handwritten, and the spell checker could not help them.
[Type V] Paraphrasing
Paraphrasing errors include examples in which a particle should have replaced a morpheme or a phrase as in (7).

(7) 행복-이라고 사람-마다 다르다.

As for happiness, every person has different ideas.

In this example, the inflectional form of the copula *ita* should have been replaced with the topic marker *은* un. In (8), a particle or the whole unit including a particle should have been replaced by a phrase or a verbal ending. This is also a paraphrasing error.

(8) 상-을 받-는 것-이 (→ 받게) 되면 기분-이 좋지.

If I come to receive an award, I feel good.

[Type VI] Spacing
The last type of learner error is spacing. Although a particle is classified as a separate word, morphosyntactically it cannot stand alone and must be attached to the preceding noun without any intervening space. In (9) the particles *에* and *서* and *는* should be stacked together without any space at all between them.

(9) 큰 도시 에서 는 때때로 길-이 혼잡해요.

In a big city, the road is very crowded.

The six error types are associated with various levels of linguistic information including lexical choice, phonology, morphology, syntax, semantics, discourse, and so on. For error classification, we do not specify the linguistic levels of particle errors. This is because an error is commonly interwoven across multiple linguistic levels in agglutinative languages like Korean. Our learner corpus with its particle error annotation was stored in XML format (see Figure 2).
PART I ****Annotated Corpus****……(omitted)………
<s><err errid="7">어머니</err> 감시는 <err errid="8">음식</err> 만들합니다.</s>
<s><err errid="9">미국음식</err> <err errid="10">한국음식</err> 만들습니다.</s>
<s><err errid="11">하고</err> 두부를 먹고, 미국 탕고기를 먹습니다.</s>
<s><err errid="12">먹시에서</err> 미국 <err errid="13">음식</err> 있습니다.</s>

PART II **** Extracted Error Rules****
<errlink relatedToError="7" type="omission" order="1" feedback="insert "회사서">
<errlink relatedToError="8" type="omission" order="1" feedback="insert 을"/>
<errlink relatedToError="9" type="omission" order="1" feedback="insert 과"/>
<errlink relatedToError="10" type="omission" order="1" feedback="insert 을"/>
<errlink relatedToError="11" type="omission" order="1" feedback="insert 을"/>
<errlink relatedToError="12" type="malformation" order="1" feedback="개시로 "에서에">
<errlink relatedToError="12" type="replacement" order="2" feedback="개시로 에에 "에에"/>
<errlink relatedToError="12" type="omission" order="3" feedback="개시로 에에"/>
<errlink relatedToError="13" type="omission" order="1" feedback="개시로 에에"/>

PART I shows the erroneous units between the markup features, <err errid="Number"> and </err>. PART II presents the extracted error rules with error types, rule ordering, and the corrected tokens.

There are several noteworthy aspects of our annotation strategy. First, it specifies both error types and the correct tokens for the corresponding erroneous units. The first nominal of the first sentence in Figure 2, which is marked as Error Rule 7, shows that there is an error in the phrase 어머니, the subject of a sentence. Error Rule 7 in PART II specifies that the relevant error type is “omission.” In addition, the error-tagging system includes a constraint of “feedback” specifying a corrected form for an error. This feature will be used for the feedback process so that learners are able to receive more specific feedback on honorific particle errors with honorific noun and verb features. In the example listed here, the feedback feature requires the user to insert the honorific nominative particle “께서” after the phrase 어머니. In order to handle particle errors efficiently, we further argue that it will be useful to add some semantic and syntactic features of honorific nouns or honorific predicates. Specifically, the honorific properties of a noun and a predicate are specified separately from the error mark-ups. This will be useful for developing automatic error detection, and particle errors related to the honorific features of a noun and a predicate can be used for feedback information.

Second, our annotation scheme allows us to specify multiple error types and represent multiple error rules associated with a single unit with rule ordering. For example, a single eojeol (space unit in Korean) like 학교-에-는-가요, hakkyo-ey-nun-kayo ‘school-to-top-go’ can be marked as an addition error (must delete 는 nun) and also as a spacing error (must put a space in front of kayo). In Figure 2, Error Rule 12 contains three types of errors in a single unit. The first is a malformation of the locative particle 에에; the wrong form 에에 needs to be corrected to 에에. The second is a replacement error; the original form 에에 needs to be replaced with the locative particle 에에. The last is an omission error; the topic marker 는 needs to be added to particle 에에. The error constraint specifies the order of rule applications with
numbers. The information on rule ordering for multiple errors will also be valuable resource for developing the autonomic processing of learner errors.

Third, when there was more than one candidate for an erroneous particle, a preference was given to a formal variant used in written Korean. The possible candidates are presented using the constraint comment. As in Error Rule 9, another possible candidate 과 is listed in the comment feature. Our annotation system allows several possible candidates for an error.

Before we move on to the results of our study, it is necessary to highlight the long distance connectivity that holds between a particle and a predicate in Korean. As already mentioned, a particle error can be associated with other errors through an argument-predicate relationship. For example, consider the following erroneous input:

(11) 부모님 이 (→께서) 나 (→저) 를 한국에 보냈어요 (→ 보내셨어요).
pwumonim-i (→kkeyse) na(→ce)-lul hankuk-ey ponaysseo (→ ponaysyessseo).
parents[+Hon]-Nom I-Acc Korea-to sent sent(Hon.)
'My parents sent me to Korea.'

In (11), the nominative particle 이 i needs to be changed into the honorific nominative 께서 kkeyse, which corresponds with the honorific subject 부모님 pwumonim 'parent.' In addition to this change, the main predicate form 보냈어요 po-nayss-eyo needs to be changed into the honorific form, 보내셨어요 po-naysyess-eyo by adding the honorific ending 시 si. Also, the object 나 needs to be changed into the humble form 저 ce 'I.' It is common for a Korean learner to make errors in both the honorific nominative particle and the predicate form. Therefore, it will be useful to incorporate linguistic information on the honorific subject particle and the corresponding predicate in the feedback constraint.

The final thing we noted in the annotation process is that there are ambiguous errors. In (12), a particle is missing after 이모 imo 'aunt.' However, it is not clear if it works as the subject or the object. For cases like (12), we chose to provide two options for the error because this will provide more accurate feedback.

(12) 이모 Ø 많이 도와주고 어머니-하고 아버지 사랑해요.
imo Ø manhi towacuko emeni-hako apeci salanghaeyo
aunt 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 love you, mother and father!”

RESULTS
In this section, we summarize our findings with respect to the basic properties of the corpus and the particle errors in the tagged corpus.

Basic Properties
Table 1 summarizes the basic properties of our corpus and lists the average number of different grammatical units such as words, paragraphs, and sentences, generated by the four different learner groups.
Table 1
Average Number of Grammatical Units in the Learner Corpus

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>Paragraphs</th>
<th>Words</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heritage</td>
<td>25</td>
<td>129</td>
<td>2,669</td>
</tr>
<tr>
<td>Nonheritage</td>
<td>25</td>
<td>188</td>
<td>1,659</td>
</tr>
<tr>
<td>Intermediate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heritage</td>
<td>25</td>
<td>261</td>
<td>2,496</td>
</tr>
<tr>
<td>Nonheritage</td>
<td>25</td>
<td>142</td>
<td>3,163</td>
</tr>
</tbody>
</table>

The average length of each of the grammatical units varies depending on learner levels and types. Beginner heritage students’ writing contains more words and sentences than that of beginner nonheritage students. For the intermediate learners, the heritage learners use more paragraphs than the nonheritage learners, which indicates higher complexity of sentence structure. However, in terms of words and number of sentences, there is little difference between these two groups. It is interesting to note that the difference in the number of sentences is larger between the beginner heritage learners and intermediate heritage learners than between the beginner nonheritage learners and intermediate nonheritage learners. When we examine the average number of grammatical units by heritage group and nonheritage group, we see that there is not much difference between the two groups other than the fact that there are higher average numbers for the heritage group.

Particle Error Distributions

This section presents our findings in terms of particle error types and patterns. In our study, the learner groups are divided into beginner heritage learners, beginner nonheritage learners, intermediate heritage learners, and intermediate nonheritage learners. Differences among the four groups are shown in Table 2 and Figure 3.

Table 2
Distribution of Particle Errors by Level and Group

<table>
<thead>
<tr>
<th>Learner Error Types</th>
<th>Beginner</th>
<th></th>
<th>Intermediate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heritage</td>
<td>Nonheritage</td>
<td>Heritage</td>
<td>Nonheritage</td>
</tr>
<tr>
<td>Omission</td>
<td>113</td>
<td>84</td>
<td>71</td>
<td>116</td>
</tr>
<tr>
<td>Replacement</td>
<td>114</td>
<td>104</td>
<td>68</td>
<td>107</td>
</tr>
<tr>
<td>Addition</td>
<td>21</td>
<td>21</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>Malformation</td>
<td>49</td>
<td>12</td>
<td>28</td>
<td>15</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Spacing</td>
<td>58</td>
<td>22</td>
<td>31</td>
<td>53</td>
</tr>
<tr>
<td>Total</td>
<td>361</td>
<td>244</td>
<td>205</td>
<td>328</td>
</tr>
</tbody>
</table>
The frequency ranking of error types by Korean learners is represented in (13).

(13) Error Type Hierarchy
Omission > Replacement > Spacing > Malformation > Addition > Paraphrasing

Figure 3 shows that the rate of omissions is similar across all four groups, while the rate of replacement errors for beginner nonheritage learners shows a spike. However, the replacement error rate for intermediate nonheritage learners is similar to that of heritage learners at both levels. At the beginner level, the most frequent particle error type is replacement for both heritage and nonheritage learners, and the omission rate takes the second highest position, close to that of replacement. In contrast, intermediate level learners show more errors in omissions than in replacements. The next most frequent error type is the spacing error, which causes more trouble for intermediate learners, as well as for heritage learners. This is due to the properties of Korean that license complex morphosyntactic constructions including compound nominals and verb + auxiliary combinations. Malformations or misspellings occupy the fourth position. Addition and paraphrasing errors are relatively low among all learners. In particular, intermediate heritage learners show the lowest rate of errors in addition and paraphrasing across different error types and in comparison with other learner groups.

It is difficult to examine whether the difference between heritage learners and non-heritage learners is statistically significant in our learner corpus because of the small size of the corpus. In addition, in order to provide an accurate analysis, it is crucial to count the error rates based upon the total number of all the particles that appeared in the given learner...
corpus. However, this task requires a much more laborious annotation process without the help of a part-of-speech tagger that can process erroneous input. Given the lack of time and staff, we have postponed a more refined analysis until the future and have focused on building up the learner corpus, developing annotation standards for particles and identifying different error patterns according to different learner groups.

In spite of the small corpus size, a notable contrast in error patterns can be found between heritage learners and nonheritage learners. The contrast suggests that the process of heritage language learning is different from that of nonheritage language learning. If this is true, then pedagogical approaches need to differ depending on learner background. Let us compare the error patterns between beginner heritage learners and beginner nonheritage learners (see Figure 4).

In general, beginner heritage learners show similar proportions of omissions and replacements, whereas beginner nonheritage learners show more replacement errors than omission errors. In terms of malformations and spacing, beginner heritage learners have a higher proportion than nonheritage beginners. Malformation errors are attributed to a mismatch between orthography and phonetic recognition in Korean. For example, the genitive particle 의 uy in Korean is pronounced [e] in the spoken language. Beginner heritage learners have greater understanding of phonological units than their ability to produce the language. It is easy for them to misrepresent the orthographical form of the genitive marker as 에 ey, which has the same phonetic realization. In addition, a particle form like 두 twu, which is synonymous with 도 to ‘also,’ used in only spoken Korean, was written as it is pronounced in heritage learners’ writing.

Now, let us consider intermediate heritage learners and intermediate nonheritage learners. As we see in Figure 5, differences between the intermediate learner groups are less
clear than those in the beginner groups. However, it is apparent that intermediate heritage learners produce more malformed particles than intermediate nonheritage learners. This phenomenon can be explained if we consider the fact that most heritage learners are likely to know more particle forms from using (or hearing) the language at home and tend to represent orthographic forms with the phonetically perceived forms. Although correct particle forms are introduced in classroom instruction, we observe in actual language classrooms that it is difficult for intermediate heritage learners to correct their orthographical errors. This observation, however, needs to be more carefully examined by tracing the same learner groups over years and determining the change in their error patterns.

Across all four groups of learners, several specific subpatterns occur within two error types. First, within omissions, the dropping rate of nominative particles is higher for heritage learners than for nonheritage learners (41% vs. 17%). Second, within replacements, beginner heritage learners overuse nominative particles in places where they should use topic particles. In contrast, beginner nonheritage learners are as likely to replace nominative particles with topic particles as they are to replace topic particles with nominative particles. Replacement errors between the nominative and accusative particles are fewer for intermediate heritage learners than for intermediate nonheritage learners (10% vs. 20%), and (incorrect) overuse of topic particles are relatively low for intermediate heritage learners compared to that of intermediate nonheritage learners (8% vs. 15%). In addition, the malformation error rate is still higher for intermediate heritage learners than intermediate nonheritage learners. These facts imply that subtypes of particle replacement differ among the heritage and nonheritage groups.

Recently there has been growing academic awareness of heritage language learning from different perspectives of linguistics including processing, acquisition, and cognition, as
well as pedagogy. Although there are more questions and issues to resolve, language learning of heritage learners has been proven to be distinct from that of nonheritage learners. In line with this, our study shows that heritage learners of Korean generate error patterns that differ from those of nonheritage learners.

SUMMARY AND FURTHER STUDY

In this paper, we explained the process of building a Korean learner corpus and adding annotation for particle error detection. We argued for the importance of learner background information in building up a learner corpus and for conducting further research with developed learner corpora. Detailed information about learners is a prerequisite for an accurate analysis and enhances the usability of the learner corpus. For example, we showed that differences exist between heritage learners and nonheritage learners with respect to particle errors. This cannot be further investigated if a corpus does not specify the language background of learners.

We also argued that error tagging needs to be linked to the feedback process. In particular, particle errors are closely associated with other components in a sentence, and this needs to be considered in the feedback process. Complex errors including spacing errors, lexical selection errors, and so on ultimately need to be incorporated into the annotation system. Rule ordering will also be needed for providing correct feedback. In addition, heritage learners seem to show unexpected errors across different levels. In school, heritage learners who are fluent in speech can bypass lower level courses and go immediately to an advanced level course. Although such learners tend to be very fluent both in speech and writing, they can generate unexpected low-level errors that are typically corrected in the low-level courses. Therefore, it will be ideal to provide more detailed information for basic level errors for heritage learners. In other words, user-oriented feedback will be more efficient than level-based feedback for heritage learners.

In terms of further study, we plan to develop more specific details of the annotation guidelines for expanding the usability of the tagged corpus. Then checking interannotator agreement will be required. We aim to expand the particle error tagging scheme to a larger size learner corpus and various Korean learners with backgrounds different from Korean American learners. Then it will be possible to elaborate error type analysis using statistical tools. While pursuing the use of open-ended user input, we also need to investigate how to use the learner corpus for other types of errors.
REFERENCES


AUTHORS’ BIODATA

Sun-Hee Lee is an assistant professor of Korean Program in the Department of East Asian Languages and Literatures at Wellesley College. Her research interests include intelligent computer-assisted language learning, syntax, semantics, corpus linguistics, and Korean linguistics.

Seok Bae Jang is a visiting assistant professor of Korean in the Department of Asian and Near Eastern languages at Brigham Young University. He is interested in intelligent computer-assisted language learning, computational linguistics, summarization evaluation, temporal and event processing, and Korean linguistics.

Sang Kyu Seo is a professor of Korean Linguistics in the Department of Korean Languages and Literature at Yonsei University. He is interested in Korean syntax, computational linguistics, and Korean lexicography.
AUTHORS’ ADDRESSES

Sun-Hee Lee
Green Hall #236B
Wellesley College
106 Central Street
Wellesly, MA 02482
Phone: 781 283 2427
Fax: 781 283 3608
Email: slee6@wellesley.edu

Seok Bae Jang
Department of Asian and Near Eastern Languages
3089 JFSB
Brigham Young University
Provo, UT 84602
Phone: 801 422 9121
Fax: 801 422 0028
Email: sbjang@gmail.com

Sang-Kyu Seo
Department of Korean Languages and Literature
Yonsei University
Seoul, 120-749, South Korea
Phone: +2 2123 2277
Fax: +2 2123 8662
Email: inaka@yonsei.ac.kr