Verb conjugation errors by learners of Korean

An entropy-based analysis of learner corpus

Chanyoung Lee
Yonsei University, South Korea

This study aims to identify patterns of verb conjugation errors that learners of Korean manifest and the factors that influence these errors through an analysis of an error-annotated learner corpus. For this purpose, a paradigmatic relations-based description of language acquisition was proposed. The predictability of each conjugation class was estimated by way of entropy, a tool for measuring predictability. Using entropy allowed us to compare the regularity of each class in detail. The results showed that there are 332 verb conjugation errors that can be classified into three types of errors: errors with vowel endings, errors with lower entropy, and errors with higher entropy. The frequency of the first two types suggests that learners make errors when producing frequently used conjugated forms and with more predictable classes. Considering this study's reproducibility and the reliability of its procedures and the results, its findings are expected to make a substantial contribution to the study of error analysis using error-annotated learner corpora.

Keywords: error-annotated learner corpus, verb conjugation, entropy, predictability

1. Introduction

Language data produced by learners have been the most important resource in the study of second/foreign language acquisition and education. However, the methods for dealing with these data are not objective but dependent on the intuition of the researcher, and the size of the data was very small in the early stages of the study. In addition, it was difficult to be representative because of the concentration on data produced in controlled environments or specific types of data. In the 1980s, as the framework of corpus linguistics was grafted into the field of

https://doi.org/10.1075/kl.20008.lee
Korean Linguistics 18:2 (2022), pp. 209–238. ISSN 0257-3784 | E-ISSN 2212-9731

second/foreign language acquisition, the learner corpus began to be created, and many of these limitations were resolved. As computerized L2 data can be accessed and processed, it has become possible to construct the language production data of a very large number of diverse learners. Since then, learner corpus research has been steadily developed technically, methodologically, and theoretically (Corder 1971, Ellis 1997, Granger et al. 2002, Lennon 1991). The learner corpus can be analyzed and processed according to various research purposes. Among them, the error-annotated learner corpus – corpora that include information about learners' production errors – is the most suitable format for language acquisition theory and educational purposes (Dagneaux et al. 1998). With it, researchers are able to analyze the various types of errors made by learners, identify the causes of their occurrence, and use them to improve educational plans.

This trend has also been found in research on Korean learners. At the beginning of the study, researches that depended on the researcher's intuition were conducted with insufficient data (Park et al. 1999, Seo 1992, Woo 1997), before the learner corpus began to be used for L2 learning and teaching research (Ko et al. 2004). The study of error-annotated Korean learner corpora began in the early 2000s (e.g., Kim 2002, Ko et al. 1999, Lee 2002, Seo et al. 2002) from a wide variety of aspects, including methodology (e.g., Kim 2002, Kim 2005, Ko et al. 2004), particles (e.g., Han 2016, Kim 2017, Lee et al. 2013, Zhang & Kang 2018), verb conjugation[1] (e.g., Lee 2019, Lee 2020), endings[2] (e.g., Han 2018), and collocation (e.g., Hong 2007).

Although learner corpus research, especially error analysis, is now actively conducted, there are points that these studies miss. First, as mentioned earlier, the research topics focus on specific grammatical forms and phenomena. Among them, in the case of verb conjugation, which is the subject of this study, are several studies that analyzed ending errors, with only a few focused on the conjugation pattern itself. Given that errors related to verb conjugation account for the second

[1] In this paper, "verb" refers to both *tongsa*, verbs, and *hyeongyongsa*, adjectives. Verbs and adjectives are distinguished in Korean's parts of speech system, but they can be misunderstood in translation. Korean *hyeongyongsa* correspond to English adjectives, but they have attributes that make them more similar to English verbs, not adjectives. English adjectives have much more in common with Korean *kwanhyeongsa*.

[2] *Emi*, meaning "ending," are elements that are affixed to verb stems and express various meanings in Korean. They are similar to suffixes, but suffixes are components of words, whereas *emi* are phonologically dependent on words. They also include phrases and clauses.
In addition, this study classified ending into the following three types: endings starting with consonant, endings starting with semi-vowels, and endings starting with vowel. For convenience of discussion, they are called consonant ending, semi-vowel ending, and vowel ending, respectively.

highest frequency of all types of errors (Section 3.2.1), it is clear that interest in this has been lacking. Second, because second/foreign language research is basically based on the existing Korean grammar system, a problem arises when this is applied uncritically. Because the native speaker and the learner have completely different ways of acquiring and using a language, it is necessary to consider the cause of the error that reflects the learner's learning pattern.

This study explores the patterns of verb conjugation errors that learners of Korean manifest and the factors that cause these errors. It was difficult to find a study on verb conjugation by analyzing large-scale error-annotated learner corpus. In addition, previous studies on learners' verb conjugation were based on the concept of regularity for Korean native speakers. To overcome this limitation, the regularity of each conjugation class is estimated by way of entropy, a tool for measuring predictability. By using this novel and concrete concept, the predictability of each conjugation class can be accurately measured. It is possible to assume a spectrum of regularity that reflects the learners' actual acquisition and to use it appropriately for analysis through this methodology. Reports in this study will have significant implications for future error analysis studies as the first entropy report assisted with the investigation of the error-annotated learner corpus.

2. Background

2.1 Analysis of Korean error-annotated learner corpora

The study of error-annotated Korean learner corpora began in the early 2000s (e.g., Kim 2002, Ko et al. 1999, Lee 2002, Seo et al. 2002). However, these early studies generally just presented the concept and need for error-annotated learner corpora (e.g., Kim 2002, Kim 2005, Wang 2003). Kim (2002) constructed an error-annotated corpus consisting of 5,465 eojeols[3] from a paper-and-pencil test administered by three Korean language education institutes, but she did not provide any information about the learners' proficiency or native language (L1). Without this information, it is difficult to conduct a proper analysis. Learner corpora mostly consist of written texts – written assignments and tests – so most research that used them focused on a limited range of subjects, such as particles and endings. However, using incorrect particles and endings are not the only types of errors learners make. They frequently make pattern-related errors, such

[3] *Eojeol* are units separated by white space in Korean. Korean is agglutinative language, so eojeol and semantic units are not coterminous. In corpus linguistics, eojeol are the basic units for the computational processing.

as syntactic constructions or morphological operations. Thus, the analysis of learner corpora for learners' error patterns must be carried out from various perspectives.

One of the most notable and recent Korean learner error annotations and analysis studies is Lee et al. (2013). They presented a methodology for constructing an error annotation corpus that analyzed particle errors in 100 learners' essays. The study is significant because they performed the analysis using a machine learning technique. Their result, Korean Learner Language Analysis (KoLLA), is available on their web page.[4] However, this study ensures the balance of the target data, but it is limited by insufficient data. If the methodology in these studies is applied to the corpus of learners in this study, more meaningful conclusions can be drawn. Meanwhile, Israel (2014) focused on automatically detecting and correcting grammatical errors in text produced by Korean language learners. This study is also based on a machine learning method while using an annotated Korean learner corpus of particle errors.

One of the most common information obtained from learner corpora is learners' error patterns. Error-annotated learner corpora show the specific types of errors that learners make, including information about particles and endings, and how each type of error occurs by proficiency level (e.g., Kim 2017, Liu 2019, Zhang 2018, Zhang & Kang 2018).

Analyses of error-annotated learner corpora have mostly focused on error patterns related to particles and endings because particles and endings have distinct grammatical functions in Korean. Particle error research has mostly been on omission, substitution, and addition, whereas ending error research has mostly been on final, connective, and pre-final endings. Final and pre-final endings express the meaning of various grammatical categories, such as tense, aspect, sentence type, mood, and honorific. Research on errors involving these ending types have examined why L2 learners of Korean find it difficult to select proper endings and possible solutions to reduce such errors.

2.2 Korean verb conjugation and regularity

Korean verb forms must change to express grammatical functions. This change in verb forms occurs by attaching special morphemes, a process known as conjugation. The central part of the conjugated form is called stem, and the parts that are attached at the end of the stem are called endings.

Endings often change their form when they are attached to stems. Some changes are full, whereas others are partial. The latter are traditionally known

[4] https://cl.indiana.edu/~kolla/

as irregular conjugations. Examples in (1) result from the final consonant rule (Lee 2013, Lee 2014) which states that the original sound pronounced before a vowel ending changes to the representative sound before the consonant ending. For example, the /s/ sound in lexeme PESTA[5] in (1a) was retained when combined with a vowel ending, but it changed to the representative sound /t/ when combined with a consonant ending. Changes to stems by the final consonant rule are universal and without exception, and such changes are known as automatic or phonetic alternation. A stem alternation's environment determines its alternation type.

(1) a. PESTA */petkko/*, */pesuni/*, */pese/*
b. NAKKTA */nakkko/*, */nakkuni/*, */nakka/*
c. ANCTA */ankko/*, */ancuni/*, */anca/*

From this point of view, this study examined the phenomenon where a verb ending with a stem drops the */u/* before endings that begin with a vowel. Example (2a) is a monosyllabic alternation, whereas (2b) and (2c) are alternations of disyllabic stems. All verbs whose stems end in */u/* are subject to this rule without exception. The standard Hangul orthography classified the dropping of */u/* as irregular conjugation. However, this study regards it as a simple dropping of a sound as described in the preceding rule and replacement of stems by consonant assimilation. Traditionally, such verbs are presented as *u*-irregular verbs, but given that speakers can only form conjugated forms if they know the rules of elimination, they are not irregular verbs.

(2) a. *ssuko~sse*, *khuci~khessta*
b. *tamkuko~tamka*, *aphuta~apha*
c. *ttaluko~ttala*, *tatalumyen~tatala*, *chiluni~chile*, *tullumyen~tullessta*

Example (3) shows examples of how the final consonant *l* in the verb is dropped under certain circumstances.

(3) a. *nolta*, *nolko*, *nolci*, *nolmyen*
b. *nonun*, *nonunya*, *non*, *nopnita*, *nosiko*

In the environment of (3b), *l* is automatically dropped, but it is maintained before the endings *-ta*, *-ko*, *-ci*, and *-myen* as in (3a). The Hangul orthography and most traditional grammar treat the dropping of *l* as irregular conjugation, but this study

[5] Capitalized words are lexemes, an abstract representation of several word forms. Certain lexemes are realized in various word forms according to various grammatical environments. In this study, verb lexemes were realized in various conjugated forms. See Haspelmath and Sims (2010) for more information on the relationship between lexemes and word forms.

considers the simple dropping of a sound. This conclusion shows that whether an alternation is predicted in the same environment is a more important criterion for determining regularity than whether an alternation occurs at all.

Verb conjugation irregularity is classified as either stem or ending irregularity (Table 1). Stem irregularity occurs when part of the stem changes when a stem and ending combine. Ending irregularity occurs when part of an ending changes. Sometimes both stems and endings may exhibit irregular conjugation.

Table 1. Types of irregular conjugation

Type	List	Example	Corresponding regular lexeme
Stem irregularity	ㅅ-irregular	NASTA 'to get well'	PESTA 'to take off'
	ㄷ-irregular	TUTTA 'to hear'	KETTA 'to roll up'
	ㅂ-irregular	TOPTA 'to help'	IPTA 'to put on (clothes)'
	르-irregular1	HULUTA 'to flow'	TTALUTA 'to follow'
	ㅜ-irregular	PHWUTA 'to scoop'	CWUTA 'to give'
Ending irregularity	ㅏ-irregular	HATA 'to do'	KATA 'to go'
	르-irregular2	PHWULUTA 'to be blue'	TTALUTA 'to follow'
Stem and ending irregularity	ㅎ-irregular	HAYAHTA 'to be white'	COHTA 'to be good'

2.3 Transition of viewpoints: Syntagmatic relations-based description and paradigmatic relations-based description

The description in Section 2.2 are based on syntagmatic relations as it describes verb conjugation as a concatenative combination of stem and ending. One of the advantages of this description is that it reveals the characteristics of Korean as an agglutinative language whose core is the concatenative combination of forms. It is also advantageous that such syntagmatic relations-based descriptions provide a systematic, categorical representation of the regularity of Korean verb conjugations. By treating regularity as alternation, conjugations can be classified as regular or irregular, achieving overall systematicity.

However, syntagmatic relations-based descriptions are limited in terms of studies on language development. What should be considered in this context is not merely grammatical skill but rather basic units that the speakers actually hear and speak. Although both stems and endings are dependent forms, learners are likely to recognize and store entire conjugated forms as individual units. Alternations are made by combining stems and endings, but a speaker may cognitively recognize such combinations as individual units without considering the various elements of the conjugated form. In short, syntagmatic relations-based descriptions do not reveal this aspect of conjugation patterns well.

Considering the shortcomings of the syntagmatic relations-based descriptions, an alternative way of understanding verb conjugation is by combining stems and endings and formalizing them as paradigmatic relations between conjugated forms that have already been combined (e.g., Chung 2015, Kawasaki 2011, Lee 2018). There are three types of endings in Korean, which are defined according to their morphophonological properties (MP): endings with consonants, endings with insert-vowels, and endings with vowels (Table 2). A conjugation paradigm is a collection of conjugated forms where individual verb lexemes are implemented according to a particular environment. Korean verb paradigms consist of conjugated forms of individual lexemes that are realized differently according to different morphophonological environments. A relevant feature of Korean paradigms is that they can be used to identify how conjugation patterns change based on their environment.

Table 2. Conjugation paradigms of Korean verbs

Verb lexeme	Ending with consonant	Ending with insert-vowel	Ending with vowel
CAPTA 'to hold'	*capko*	*capuni*	*capa*
IPTA 'to put on (clothes)'	*ipko*	*ipuni*	*ipe*
TOPTA 'to help'	*topko*	*towuni*	*towa*
TEPTA 'to be hot'	*tepko*	*tewuni*	*tewe*

This description of verb conjugation differs greatly from the aforementioned syntagmatic relations-based descriptions. In this type of description, stem and ending units do not have independent status and individual conjugated forms form a paradigmatic set at the same level. Using paradigmatic relations-based descriptions does not require the presentation of a different paradigm for each lexeme as shown in Table 1 because each lexeme can be abstracted and classified to some degree.

Table 3 shows the abstracted common parts created from the conjugation paradigm of the four lexemes – CAPTA, IPTA, TOPTA, and TEPTA – presented in Table 2. They are classified into two types according to the commonality of the conjugation pattern. Each pair has the same conjugation pattern except for vowel harmony. The lexemes in these two categories have the same conjugation patterns in environments with consonant endings but different conjugation patterns in the other two environments. In this regard, they are divided into "ㅂ-class 1" and "ㅂ-class 2," respectively.

Table 3. Conjugation class and scheme

Verb lexeme	Ending with consonant	Ending with insert-vowel	Ending with vowel
CAPTA 'to hold'	*capko*	*capuni*	*capa*
IPTA 'to put on (clothes)'	*ipko*	*ipuni*	*ipe*
ㅂ-class 1	V*p-kko*	V-*puni*	V-*pa/pe*
TOPTA 'to help'	*topko*	*towuni*	*towa*
TEPTA 'to be hot'	*tepko*	*tewuni*	*tewe*
ㅂ-class 2	V*p-kko*	V-*wuni*	V-*wa/we*

Learners who have learned Korean verb conjugation patterns will be familiar with the Korean verb conjugation paradigm, though it may be somewhat incorrect. What patterns do learners think are regular or irregular in the paradigm in Table 4? If the lexeme's consonant-ending form is *a*, then it can be class 1–4, but if it is *b*, then it must be class 5. The only possible candidate is much more predictable than the four candidates (in fact, prediction is unnecessary if the only candidate is possible). Thus, classes with consonant-ending form *a* are less predictable than those with form *b* and thus more irregular.

Table 4. Artificial verb conjugation paradigm

Conjugation class	Ending with consonant	Ending with insert-vowel	Ending with vowel
Class 1	*a*	*c*	*g*
Class 2	*a*	*c*	*h*
Class 3	*a*	*d*	*h*
Class 4	*a*	*c*	*i*
Class 5	*b*	*f*	*j*

Note. The elements (*a*, *b*, *c*, ...) in each column represent conjugation patterns rather than a form in a strict sense.

A comparison of classes 2 and 3 shows that they both have *a* as their consonant-ending form and *h* as their vowel-ending form, so they would be expected to have similar levels of regularity. However, their regularity differs in their insert-vowel ending forms because *c* in class 2 overlaps with classes 1 and 4 throughout the paradigm, whereas *d* in class 3 is unique. In other words, the conjugated forms that constitute class 3 are easier to predict than those in class 2, so class 3 is more regular.

For example, the two *c*s in classes 1 and 2 have different forms but the same conjugation pattern. If the conjugation pattern of a conjugation class is so unique that it can be quickly and easily determined that only a few of them are in the class, the probability of choosing the correct form is high, giving it high predictability. This situation is the basic idea behind regularity based on paradigmatic relations. Thus, the regularity of conjugations in paradigmatic relations-based descriptions is defined as the predictability of conjugation class and individual conjugated forms. From this perspective, regularity is not a concept that is dichotomously divided into 'irregular' and 'regular', but a continuous one. In other words, Korean conjugation classes can be compared according to their relative degrees of regularity. This methodology is discussed in detail in Section 3.1.

3. Methods

This study analyzed large-scale conjugated form errors using Korean error-annotated learner corpus to determine whether there is a constant tendency for error occurrence and to identify what factors affect the regularity with which learners produce each conjugation class. Paradigmatic relations-based descriptions were used to emphasize the two aspects to be revised from the existing descriptions. First, stems and endings should not be segmented as they create a conjugated form as the basic unit of production. Second, the regularity of each conjugation class should be judged by its predictability of producing a conjugated form, not by an alternation involving stems and endings. The predictability is continuous so that each class's predictability can be quantified for comparison. To achieve this, entropy was used to assess the regularity of conjugation classes.

3.1 Regularity measurement using entropy

Entropy, a concept that originated in the natural sciences, is the unusable energy generated in the process of energy conversion. Shannon (1948, 1951) applied the concept to information theory by substituting it for informational concepts. Entropy in information theory is used to express the amount of information in an event that is closely related to the probability of each event occurring. The probability of an event occurring is inversely proportional to the amount of information provided such that the higher the probability of an event, the lower the amount of information in that event.

A unit of entropy, known as "bit", is defined as $-log_{(2)}P(x)$ based on the probability of each event occurring within a specific range. If the probability of two events occurring is 0.5 each, like in a coin flipping situation, then the entropy is 1 bit. As such, the uncertainty surrounding the probability of an event occurring can be accurately quantified through entropy, and these values can be meaningfully compared. This use of entropy in information theory attracted the attention of linguists who want to quantify the complexity in inflection classes (e.g., Ackerman et al. 2009, Stump & Finkel 2013). Ackerman et al. (2009, p. 63) and Stump & Finkel (2013, p. 296) defined entropy for set X, *H(X)*, as the weighted average of the probability of each event occurring in order to determine the complexity of the set through in terms of relationships between the forms in the paradigm.

(4) Entropy

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \times 100$$

Entropy may not simply represent the predictability of an entity or set. Like conditional probability, which is the probability that event A will occur when event B occurs, entropy can also be expressed in terms of conditional entropy of element A that is affected by the entropy of element B. The conditional entropy of A with respect to B, *H(A|B)*, is given by the following:

(5) Conditional entropy

$$H(A|B) = -\sum_{y \in B} P(y) \sum_{x \in A} P(x|y) log_2 P(x|y) \times 100$$

If information about B can be used to predict A, then the conditional entropy *H(A|B)* will be lower than *H(A)*. If information about B cannot be used to predict A, then the conditional entropy *H(A|B)* will be the same as *H(A)*. As such, conditional entropy reveals the existence of a relationship between two elements or sets as well as the direction and degree of correlation of that relationship.

In Korean, each conjugation class forms a paradigm of conjugated forms realized according to their MPs. Therefore, methods for generating the conditional entropy of a conjugation class must consider all conjugated forms that constitute its paradigm. The conditional entropy of the inflection class presented by Stump & Finkel (2013, pp. 311–313) is modified as follows for use with Korean verb conjugation classes. Suppose that one of the MPs constituting conjugation class *I* is *m* and its corresponding conjugated form *e* and assume that a paradigm consisting only of conjugation classes capable of having the same conjugated form as *e* for *m*'s MP is a reduced paradigm, *R*. The conditional entropy of the remaining MP for *m*, *m′* is obtained, and the average of these is the MP entropy of the conjugation class (6).

(6) MP entropy of the conjugation class

$$H_{i,m} = \frac{\sum_{m' \neq m} H_R\,(m'|m)}{|MP| - 1}$$

In this way, the conjugation class entropy (CCE) for conjugation class *I* is obtained from the entropies of the three MPs as in (7).

(7) Conjugation class entropy

$$CCE_m = \frac{\sum_m H_{i,m}}{|MPS|}$$

To calculate the CCE using the example "ㅂ-class1," the conjugation paradigm "CAPTA," the lexeme corresponding to "ㅂ-class1" is {capkko, capuni, capa}. The reduced paradigm *R* consisting of conjugation classes that may have the same form as each conjugated form can be assumed as follows (Table 5–7):

Table 5. Reduced paradigm R for conjugated form with consonant ending of lexeme "CAPTA"

Conjugation class	Ending with consonant	Ending with insert-vowel	Ending with vowel
ㅂ-class1	capkko	capuni	capa
ㅂ-class2	capkko	cawuni	cawa
ㅂ-class3-A	capkko	cani	cae
ㅍ-class	capkko	caphuni	capha
ㄿ -class	capkko	calphuni	calpha
ㅄ -class	capkko	capssuni	capssa
ㄼ -class-B	capkko	calpuni	calpa
MP entropy	280.74		

Table 6. Reduced paradigm R for conjugated form with insert-vowel ending of lexeme "CAPTA"

Conjugation class	Ending with consonant	Ending with insert-vowel	Ending with vowel
ㅂ-class1	capkko	capuni	capa
ㄹ-class	capulko	capuni	capule
ㅡ-class	capuko	capuni	capa
ㅂ-class3-A	capupkko	capuni	capue
MP entropy		175	

Table 7. Reduced paradigm R for conjugated form with vowel ending of lexeme "CAPTA"

Conjugation class	Ending with consonant	Ending with insert-vowel	Ending with vowel
ㅂ-class1	capkko	capuni	capa
ㅡ-class	capuko	capuni	capa
ㅏ-class1	capako	capani	capa
MP entropy			125.16

These conjugation classes, which are each selected around conjugated forms "capkko," "capuni," and "capa," are a collection of classes whose predictability decreases because of these conjugated forms. The average of these values is 186.22. This value is the MP entropy value only for lexeme "CAPTA". In this paper, the conditional entropy value of all Korean lexemes is calculated, and the weighted average value according to the number of lexemes is treated as the entropy value of the corresponding conjugation class. This value is the CCE. On the basis of the result of the calculation, the MP entropy value of the entire ㅂ-class is 281.87.

The CCE was applied to a specific conjugation class to determine the amount of entropy values by which the predictability of another conjugation class changes. Then the contribution of the conjugation class to the change in predictability of the entire conjugation system was generated by adding these values. First, the list of the members of conjugation class *j* that is realized in the same conjugated form as conjugated form *e* corresponding to MP *m* of conjugation class *i* is determined. The CCE of conjugation class *j* was then obtained by subtracting the calculated CCE from conjugation class *i*. The sum of these results is the prediction entropy value for MP *m* in conjugation class *i*. Similarly, the predictive entropy of conjugation class *i* was calculated in the same way for the other two MPs, and the average of the three was taken. This average value was used to determine the degree of regularity for each conjugation class, also known as its conjugation-class predictive entropy (CCPE) (8).

(8) Conjugation-class predictive entropy

$$CCPE_i = \frac{\sum_m \sum_j CCE_{j-i,m}}{|MPS|}$$

To calculate the CCPE using the example "르-class1," the value obtained by subtracting the CCE calculated excluding "르-class1" from the CCEs of other conjugation classes that may be realized in the same way as the conjugated form constituting the conjugation paradigm of "르-class1" is as follows:

Table 8. Reduced paradigm of "르-class1" and differences in entropy

Ending with consonant	Difference	Ending with insert-vowel	Difference	Ending with vowel	Difference
르-class2	29.25	ㄷ-class2	12.98	ㅡ-class	1.06
ㅡ-class	29.25	ㅀ-class	27.27	ㅏ-class	1.50
		ㅡ-class	1.32		
		르-class2	12.98		
Sum	**58.50**		**54.55**		**2.56**

The average of the three entropy differences shown in Table 8 is 38.54, which is the CCPE of "르-class1". This procedure produced a spectrum of verb conjugation regularity for all the conjugation classes. The list of predictive entropy for all verb conjugation classes in Korean is presented in Appendix A (adapted from Lee 2018).

3.2 Analysis of conjugation errors in learner corpus

3.2.1 *Information about the error-annotated corpus*

An error-annotated learner corpus was used to analyze conjugation errors by L2 learners of Korean. The error-annotated learner corpus referred to here is part of a large-scale learner corpus called the Korean Learner Corpus (National Institute of Korean Language 2018). The Korean Learner Corpus was developed under the supervision of the National Institute of Korean Language which develops and implements language policy and conducts linguistic research. The development of this corpus began in 2015 and is still ongoing. The portion of the corpus used in this study was developed between May 2015 and November 2018. In order to overcome the limitations of the scale, balance, and diversity of previous studies on learner corpora, the Korean Learner Corpus was constructed of Korean materials produced by materials from learners with as much nationality and L1 diversity as possible. As of 2018, a total of 2.6 million eojeols made up the raw corpus, of which 2 million were written and 580,000 were oral. These eojeols were produced by learners of 80 L1s from 124 countries. Overall information on the Korean Learner Corpus is presented in Appendix 2.

The Korean Learner Corpus is composed of three types of corpus: raw corpus, morpheme-annotated corpus, and error-annotated corpus (Table 9).

The error-annotated corpus used in this study consisted of representations of the original forms and types of errors learners made. Error-annotated corpora need to indicate the type of errors that learners made to understand patterns in

Table 9. Frequency of corpus by type

	Written		Spoken	
	Token	**Sample**	**Token**	**Sample**
Raw corpus	2,021,991	15,983	579,391	1,251
Morpheme-annotated corpus	1,509,990	12,904	381,029	820
Error-annotated corpus	348,532	3,093	218,694	507

their type or frequency and how those patterns change as learners' proficiency improve. Among these, error type annotations describe the division of grammatical components, categories, and units to which the error belongs. This information is relevant to the conjugation errors analyzed in this study. There were 33,294 such annotations (Table 10), of which phoneme errors were the most common and conjugation errors were the second-most common. The fact that conjugation errors, the topic of interest in this study and a topic that has been explored in detail, accounted for >10% of production errors indicates that the learners are struggling to properly conjugate verbs.

Table 10. Frequency and proportion by type of errors

Type	**Frequency**	**Proportion (%)**
Phoneme	14,458	43.43
Conjugation	3,964	11.91
Tense	3,199	9.61
Register	2,561	7.69
Honorific	1,567	4.71
Phonological rule	1,466	4.40
Others	6,079	18.25
Total	**33,294**	**100**

Note. "Others" includes tense (3,199 cases; 9.61%), written/spoken (2,561 cases; 7.69%), honorific (1,567 cases; 4.71%), and 14 more types.

3.2.2 *Extracting paradigm predictive errors*

The National Institute of Korean Language made the 2015–2018 Korean Learner Corpus available on the web. It can be searched according to conditions, such as location, pattern, or error type as described earlier. Search results can be downloaded as Microsoft Excel files. These functions were used and analyses were conducted as follows. First, the Korean Learners' Corpus search engine was accessed,[6] and the "error-annotated corpus search" tab was clicked on. Under error types, "inflection (conjugation)" was selected, and the search button was clicked. This search produced 2,365 results. Then the download button at the bottom of the screen was clicked to download the file in an Excel worksheet format. The Excel file provided 15 types of information about each error.[7]

The focus of this study is paradigm predictive errors, so the following post-processing was conducted. First, errors related to paradigm predictability were extracted from the full set of conjugation errors. This extraction was done manually because this information was not annotated. A total of 332 sentences were extracted. The types of errors excluded from the whole can be classified as follows:

Table 11. Types of errors excluded from the analysis

Type	Frequency	Example
Ending usage errors	1,521	*kekcenghayssnunta* (√*kekcenghayssta*) (to worry) *tayanghantanun* (√*tayanghatanun*) (to vary) *pwasupnita* (√*pwasssupnita*) (to see)
Errors caused by using copula *ita*	142	nanun cikum *haksaynginta* (√*haksayngita*) I-TOP now student-COP *hunceklako* (√*huncekilako*) (to be a trace)
Basic form errors	128	cwungkwuk salamto *phohamtoyta* (√*phohamtoynta*) china people-FOC be included eluni *toyse* (√*twayse*) adult-SBJ to become
Lexeme/stem usage errors	125	*sakweta* (√*sakwinta*) (to make friends) pwumonimkkey unhyeylul parents-DAT kindness-OBJ *kiphtanun* (√*kaphnuntanun*) to be deep (√to repay)

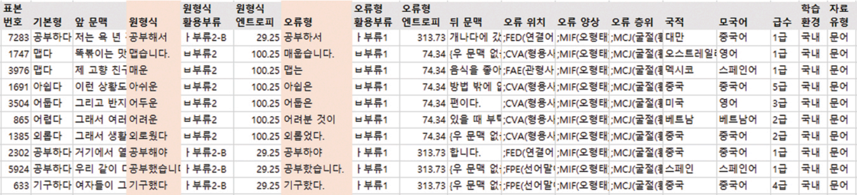
[6] https://kcorpus.korean.go.kr/index/goMain.do

[7] The Microsoft Excel file included the sample's number, its error form, the POS of the error, the original form, the preceding context, keywords, the succeeding context, the location of the error, the pattern of the error, the type of the error, the nationality of the learner, the learner's L1 language, the learner's Korean level, whether the learning environment was in Korea or not, and whether the data were originally written or spoken.

Table 11. *(continued)*

Type	Frequency	Example
Indistinguishable errors	108	wulinun hangsang toni *cweta* (√???) we-TOP always money-SBJ kiswuksaeyse *naokinta* (√???) dormitory-LOC
Correct forms	9	

To generate predictive entropy values, it is necessary to indicate which conjugation classes the errors are in and how high or low the entropy values are for the conjugation class that the error belongs. The Excel file showed the base form of the conjugated form of the error, so the conjugation class to which each conjugated form belonged, and their corresponding entropy values were entered. Part of the data format after processing is shown in Figure 1.



표본 번호	기본형	앞 문맥	원형식	원형식 활용부류	원형식 엔트로피	오류형	오류형 활용부류	오류형 엔트로피	뒤 문맥	오류 위치	오류 양상	오류 층위	국적	모국어	급수	학습 환경	자료 유형
7283	공부하다	저는 욕 년	공부해서	ㅏ부류2-B	29.25	공부하서	ㅏ부류1	313.73	개나다에 갔	;FED(연결어	;MIF(오형태	;MCJ(굴절(	대만	중국어	1급	국내	문어
1747	맵다	똑볶이는 맛	맵습니다.	ㅂ부류2	100.25	매웁습니다.	ㅂ부류1	74.34	{우 문맥 없	;CVA(형용사	;MIF(오형태	;MCJ(굴절(	오스트레일	영어	1급	국내	문어
3976	맵다	제 고향 친	매운	ㅂ부류2	100.25	맵는	ㅂ부류1	74.34	음식을 좋아	;FAE(관형사	;MIF(오형태	;MCJ(굴절(	멕시코	스페인어	1급	국내	문어
1691	아쉽다	이런 상황도	아쉬운	ㅂ부류2	100.25	아쉽은	ㅂ부류1	74.34	방법 밖에	;CVA(형용사	;MIF(오형태	;MCJ(굴절(	중국	중국어	5급	국내	문어
3504	어둡다	그리고 반지	어두운	ㅂ부류2	100.25	어둡은	ㅂ부류1	74.34	편이다.	;CVA(형용사	;MIF(오형태	;MCJ(굴절(	미국	영어	3급	국내	문어
865	어렵다	그래서 여러	어려운	ㅂ부류2	100.25	어려분 것이	ㅂ부류1	74.34	있을 때 부	;CVA(형용사	;MIF(오형태	;MCJ(굴절(	베트남	베트남어	2급	국내	문어
1385	외롭다	그래서 생활	외로웠다	ㅂ부류2	100.25	외롭었다.	ㅂ부류1	74.34	{우 문맥 없	;CVA(형용사	;MIF(오형태	;MCJ(굴절(	중국	중국어	2급	국내	문어
2302	공부하다	거기에서 열	공부해야	ㅏ부류2-B	29.25	공부하야	ㅏ부류1	313.73	합니다.	;FED(연결어	;MIF(오형태	;MCJ(굴절(	중국	중국어	1급	국내	문어
5924	공부하다	우리 같이	공부했습니	ㅏ부류2-B	29.25	공부핬습니다.	ㅏ부류1	313.73	{우 문맥 없	;FPE(선어말	;MIF(오형태	;MCJ(굴절(	스페인	스페인어	1급	국내	문어
633	기구하다	여자들이 그	기구했다	ㅏ부류2-B	29.25	기구핬다.	ㅏ부류1	313.73	{우 문맥 없	;FPE(선어말	;MIF(오형태	;MCJ(굴절(	중국	중국어	4급	국내	문어

Figure 1. Post-processed error data for analysis

To measure the reliability of the annotation, two experienced annotators classified the types of errors on our sub-corpus. The types presented in Table 11 are not subject to this study, so the annotators annotated only 332 errors. Cohen's kappa, one of the well-known indexes, was used to measure agreement (Landis & Koch 1977). The agreement rate was 96.73% for the error type (Cohen's kappa = 0.949). The high values can be explained by the fact that our annotators were highly trained and the entropy values, which is the standard of annotation, were objective.

4. Results and discussion

4.1 Analysis by type

The 332 cases with conjugation class-related verb conjugation errors were divided into errors with vowel ending (154 cases), errors with lower entropy (132 cases), and errors with higher entropy (46 cases).

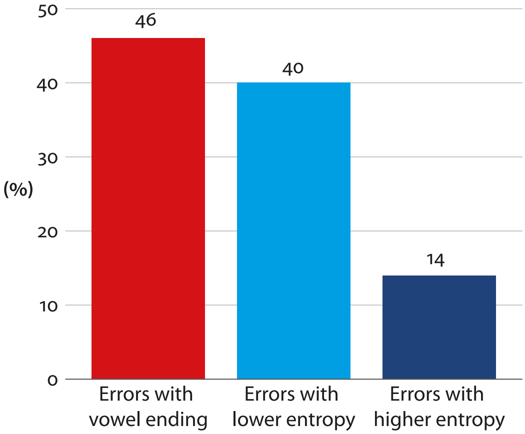


Figure 2. Ratio by error type

As in Figure 2, "Errors with vowel ending" (Type 1) accounts for the highest frequency (154 errors) of all errors. Although Type 1 is not directly related to the relative difference in entropy, it is included because it is closely related to the regularity recognition of learners about conjugation. As mentioned earlier, regularity is closely related to predictability. Regularity of conjugation can be defined as predictability of conjugation class and individual conjugated form. Among them, entropy is a measure of the predictability of each conjugation class within the entire paradigm. On the other hand, differences in regularity according to predictability can be found not only among conjugation classes but also among conjugated forms within one conjugation class. If learners' errors are regarded as a kind of language variation, there is a phenomenon that leads to a more predictable form, or analogy (Hopper & Trougott 2003). The most important factor affecting the analogy phenomenon is frequency, and vowel ending form occupies the highest frequency among conjugated forms constituting conjugation class. Considering this point of view, learners tend to be attracted to the vowel ending form, which is a form with high predictability, when they do not know the correct conjugated form of a verb lexeme. Since these types of errors appear very frequently, errors with vowel ending can be handled in conjunction with other types related to entropy.

"Errors with lower entropy" (Type 2) is a type that produces the conjugated form belonging to the conjugation class that has a lower entropy value than the conjugation class to which the correct form belongs. If the entropy value of the conjugation class to which the lexeme of the error form belongs is lower than the entropy value of the conjugation class to which the lexeme of the original form belongs, this error belongs to the "Errors with lower entropy" type. "Errors with higher entropy" (Type 3) is a type that produces a conjugated form belonging to a conjugation class that has a higher entropy value than the conjugation class to which the correct conjugated form belongs. If the entropy value of the conjugation class to which the lexeme of the error form belongs is higher than the entropy value of the conjugation class to which the lexeme of the original form belongs, this error belongs to the "Errors with lower entropy" type. That is, errors that include the use of more regular conjugated form for learners belong to Type 2, whereas Type 3 includes errors that involve the use of more irregular conjugated form.

Regarding errors with vowel ending (Table 12), consonant or insert-vowel endings are correct, but errors appeared when creating the vowel ending form. This result is supported by lexeme conjugation paradigms.

Table 12. Examples of errors with vowel ending

Error form	Original form	Lexeme	Paradigm
kukes-ul makalyeko	*makulyeko*	MAKTA 'to block'	*makko*, *makuni*, ***maka***
phikonhaci anhamyen	*anhumyen*	ANHTA 'to be not'	*anhko*, *anhuni*, ***anha***
hankwuke cal moshaynikka	*moshanikka*	MOSHATA 'cannot'	*moshako*, *moshani*, ***moshay***
chinkwuka cohapnita	*cohsupnita*	COHTA 'to be good'	*cohko*, *cohuni*, ***coha***
kathi salanun	*sanun*	SALTA 'to live'	*salko*, *sani*, ***sala***

Verb lexemes form a conjugation class consisting of three classes of conjugated forms: conjugated forms with consonants, insert-vowel, and vowel endings. There is a significant difference in the frequency with which each type is used. In Lim (2004), the change in conjugation paradigm is generally made on the basis of conjugated forms with vowel ending, which is attributed to the high frequency of use. Vowel endings are used much more frequently than consonant and insert-vowel endings because most Korean endings begin with vowels, especially high-frequency endings such as "*-e, -ess-*". As mentioned earlier, speakers tend to be drawn to more predictable and therefore more regular forms. This phenomenon, which is called analogy, often occurs even in learners who have not fully learned the correct conjugated form. The most frequent form in one conjugation class is the vowel ending form, so the phenomenon shown in Table 12 occurs frequently.

There were 132 errors with lower entropy, which reflects greater predictability. These errors accounted for 40% of all errors (Table 13).

Table 13. Examples of errors with lower entropy

Error form	Original form	Original conjugation class (entropy)	Error conjugation class (entropy)
yeylul tulumyen	*tulmyen*	ㄹ-class [*tulko*, *tuni*, *tule*] (1049.30)	ㄷ-class2 [*tutko*, *tuluni*, *tule*] (72.48)
himtunun kesi manhta	*himtun*	ㄹ-class [*himtulko*, *himtuni*, *himtule*] (1049.30)	ㅡ-class [*himtuko*, *himtuni*, *himte*] (412.43)
pwulkokika maywuciman	*maypciman*	ㅂ-class2 [*maypko*, *maywuni*, *maywe*] (100.25)	ㅜ-class1-B [*maywuko*, *maywuni*, *maywe*] (3.41)
hakkyoka kakkapunikka	*kakkawunikka*	ㅂ-class2 [*kakkapko*, *kakkawuni*, *kakkawe*]	ㅂ-class1 [*kakkapko*, *kakkapuni*, *kakkape*] (94.34)

This type of error pattern is identified by comparing the entropy values of the conjugation classes of the correct and error forms. Table 13 shows the entropy values of each class in parentheses and the conjugation classes corresponding to the correct and error forms. As seen from the entropy values in parentheses, all errors in this type produced a conjugated form belonging to a conjugation class with a lower entropy value than the original conjugation class. Given the correlation between entropy values and predictability discussed in Section 3, this phenomenon does not help learners to produce correct conjugated forms, but it does help forms be more predictable within the overall paradigm despite being incorrect. In terms of paradigmatic relations-based descriptions, this error pattern fits this study's hypothesis that speakers are more likely to make predictable conjugated forms.

Two factors appear to influence this phenomenon. First, as learners learn grammar items and irregular verbs step by step, the Korean verb conjugation paradigm rarely operates within the learning process. Therefore, learners have to choose the correct conjugated form based on an incomplete paradigm, which leads to errors. Second, learners who are unsure of the exact conjugated form will tend towards making more efficient and easier decisions, thereby reducing confusion with other potential candidates.

There were 46 errors with higher entropy, accounting for about 14% of the total number of errors (Table 14).

Table 14. Examples of errors with higher entropy

Error form	Original form	Original conjugation class (entropy)	Error conjugation class (entropy)	Confusion-trigger verb
nolaylul tulko	*tutko*	ㄷ-class2 [*tutko, tuluni, tule*] (72.48)	ㄹ-class [*tulko, tuluni, tule*] (1049.30)	TULTA 'to hold'
nolaylul pwulessta	*pwullessta*	르-class1 [*pwuluko, pwuluni, pwulle*] (38.54)	ㄹ-class [*pwulko, pwuluni, pwule*] (1049.30)	PWULTA 'to blow'
cipeyse swiwumyen	*swimyen*	ㅟ-class-A [*swiko, swini, swie*] (0.00)	ㅂ-class2 [*swipko, swiwuni, swiwe*] (100.25)	SWIPTA 'to be easy'

Entropy reflects the fact that correct conjugated forms belonging to more predictable conjugation classes are more likely to be created than errors belonging to less predictable classes. Thus, this type of error can be interpreted as not supporting this study's hypotheses relating to the concept of paradigmatic relations-based description and regularity. However, this hypothesis does not threaten this study's hypothesis that the frequency is relatively low compared with that of the first and second types. The difference between Type 3 and other types is more meaningful when classified by level rather than the total frequency, which will be examined in Section 4.2.

Another layer of factors may have contributed to the occurrence of this type of error. Whether this is the case can be determined by the confusion-trigger verb in Table 14. In some cases, the predictability of the conjugation class to which the lexeme belongs may affect the formation of the error form. However, sometimes one lexeme may be confused with another, causing learners to generate errors by producing conjugated forms that correspond with other lexemes.

4.2 Analysis by level

The pattern of conjugation errors was analyzed according to the level of learning. It was hypothesized that the higher the level, the lower the frequency of overall types of errors. This hypothesis was partially supported (Table 15).

As the learner's proficiency increases, the number of errors decreases, which is supported by this study's hypothesis. However, interesting patterns emerged when we took error types into account. First, the frequency of Type 1 errors decreased as learner proficiency increased as with overall error frequency, except at level 6. However, the relative frequency of Type 2 errors, those with lower entropy, and Type 3 errors, those with higher entropy, were slightly different. The most noticeable characteristic of Type 2 is that it shows a generally homogeneous

Table 15. Relative frequency of error types by proficiency level

Level	Frequency of error	Relative Frequency of error	Type 1 (#)	Type 2 (#)	Type 3 (#)
1	85	121.40	77.12	41.42	2.86
2	70	106.05	54.54	31.81	19.69
3	74	128.60	50.40	52.13	26.07
4	50	90.96	36.38	43.66	10.91
5	29	56.26	9.70	34.92	11.64
6	24	37.44	15.60	15.60	6.24

Note. Type 1 = errors with vowel ending, Type 2 = errors with lower entropy, Type 3 = errors with higher entropy.

frequency with little differences except at the sixth level, and the change in the frequency of errors according to the level is not significant. On the other hand, it is characterized by low frequency at the beginner and advanced levels, whereas the highest error frequency in the level 2 and level 3 is in Type 3.

How can these changes in relative frequency be understood? Type 2 errors fit into the concept of paradigmatic relations-based descriptions and regularity, whereas Type 3 errors do not as they occurred because of individual lexical confusion. Because Type 2 is a universal tendency that appears according to the learner's predictability for conjugated form, the consistently high error frequency can be understood until the advanced level of overall understanding of the verb conjugation paradigm. Meanwhile, as mentioned earlier, many of the Type 3 errors were due to the existence of confusion-trigger verbs. They occur at a stage where the acquisition of the verb form, meaning, and usage has not been properly achieved. Considering this, learners do not make many Type 3 errors because only few vocabulary words are learned at the beginner level. As the level goes up, the amount of vocabulary learning increases, and confusion between verbs occurs because of incomplete learning, which is the primary cause of such errors. When learners enter the advanced stage, they understand the differences between vocabulary words better, and the frequency of errors becomes lower.

5. Conclusion: Implications of the use of Korean corpora for developmental research on Korean

This study analyzed verb conjugation errors generated by learners using the Korean Learner Corpus. First, we pointed out the problem with the existing syntagmatic relations-based description and proposed a paradigmatic relations-based description as an alternative. On the basis of this perspective, the regularity of all Korean verb conjugation classes was measured using entropy as a measuring tool. 332 errors that relate to conjugation class predictability from the error-annotated corpus of Korean Learner Corpus were extracted, analyzed, and classified into three: errors with vowel ending, errors with lower entropy, and errors with higher entropy. The "errors with vowel ending" and "errors with lower entropy" types had high frequencies, suggesting that learners tend to produce more predictable and more regular conjugated forms than the original forms. In the analysis based on proficiency level, it was found that predictability-based errors were steadily made until learners reliably acquired the Korean verb conjugation paradigm, whereas errors caused by the interference of confusion-trigger verbs appeared mainly in the intermediate level.

Error analyses have been conducted on learner corpora, but they were limited. First, the corpora were often private, not generated systematically, small, inconsistent, and did not use clear or consistent error annotation criteria. Such corpora were not cross-checked, and there were some bias and error in relation to the intuitive annotations. These variables had significant influence on the results, so they should have been controlled for. The Korean Learner Corpus was used in this study to overcome these limitations. Using this corpus avoided many of the aforementioned limitations, including data quantity, the balance of L1 and nationalities, and the consistency of annotation. The biggest advantage of this large-scale corpus was that various studies were performed using it, making their results comparable and promoting reproducibility.

Many studies have been conducted on how to present and teach verb conjugations. However, verb conjugation regularity based on L1 Korean speakers is traditionally presented to learners. From this perspective, even if learners' errors are extracted from a corpus, analyses will still not be correct. This study proposed an alternative analysis method based on paradigmatic relations-based descriptions and entropy. Verbs that were considered irregular in the past were not so for learners. The verbs were considered regular on L1 basis and belonged to classes that make it difficult to predict the appropriate conjugated forms for learners. The entropy-based analysis revealed that learners tend to make more errors that relate to conjugated forms belonging to classes with lower entropy values than correct classes. Therefore, there is a need to focus more on the conjugation of verbs that belong to classes that are difficult for learners to predict. That is, irregular verbs.

This study's findings show the "reason" learners make conjugation errors and the results of this study can be applied to practical pedagogical fields. In current Korean language textbooks, only explanations of individual ending items and irregular conjugation are presented. It is difficult to see that the difficulties experienced by learners in creating conjugated forms are actively reflected. In consideration of this, it is necessary to present grammar items and vocabularies that are useful for learning Korean. At the same time, learners' use needs and difficulties in learning should be considered when constructing textbooks. Also, there is a tendency in textbooks to thoroughly separate stem and ending by adopting Korean grammar description methods. There are advantages for learners when they recognize them separately, but there are also advantages when they accept the entire conjugated form as a unit. Therefore, rather than treating the conjugated form as being completely separated into stems and endings, it can be a way to train them to operate as a unit within a sentence or discourse.

Although this study is significant because it is the first to analyze learners' language generation based on a large-scale error-annotated learner corpus, it had some limitations. First, even if the error annotations were made according to unified guidelines, there were still inconsistencies in the annotation results, so the data had to be further modified. Automatic modification is not a viable solution because learners make a wide variety of unexpected errors. As the amount of error-annotated corpus accumulates, future research can automate error annotation, allowing for more consistent analysis.

In addition, it is important to consider the learners' L1 in relation to the types of errors they make when carrying out a learner error analysis. On the basis of the interlanguage concept (Selinker 1972), there will be differences in the patterns of errors that occur in learning Korean depending on the learner's L1. This study did not focus on this element, but the Korean Learner Corpus contains information about learners' L1. Subsequent research will benefit from including this information in their analyses.

Research using learner corpora have become increasingly common. However, little meta-research about how best to use such corpora have been conducted. Given that the core value of studying corpora is the reproducibility of such studies, it is hoped that the processes and results of this study will contribute to future research.

References

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In *Analogy of Grammar: Form and Acquisition* ed by James P. Blevins & Juliette Blevins, 54–82. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199547548.003.0003

Chung, Kyeongjae. 2015. Historical change of the Korean conjugation system. Ph.D. dissertation. Korea University.

Corder, Stephen P. 1971. Idiosyncratic dialects and error analysis. *Internation Review of Applied Linguistics* 9: 147–159. https://doi.org/10.1515/iral.1971.9.2.147

Dagneaux, Estelle, Sharon Denness & Sylviane Granger. 1998. Computer-aided error analysis. *System* 26.2: 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

Ellis, Rod. 1997. *Second Language Acquisition*. Oxford/New York: Oxford University Press.

Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds). 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/lllt.6

Han, Songhwa. 2016. A study on the errors and use of the particle 'un/nun' by Korean learners: Focusing on their Korean proficiency and their native languages. *Eomunlonchong* 70: 111–151.

Han, Songhwa. 2018. The study of the usage and errors of Korean final endings. *Grammar education* 33: 165–210. https://doi.org/10.21850/kge.2018.33..165

Haspelmath, Martin. & Andrea D. Sims. 2010. *Understanding morphology (2nd ed.)*. London: Hoddor Education.

Hong, Hye Ran. 2007. Error analysis in grammatical collocation of KFL Learners: Focusing on learner's corpus and advanced Korean learners' composition. *Cross-cultural studies* 11.1: 23–52. https://doi.org/10.21049/ccs.2007.11.1.23

Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization* (2nd ed.). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Israel, Ross. 2014. Building a Korean particle error detection system from the ground up. Ph.D. Dissertation, Indiana University.

Kawasaki, Keigo. 2011. 'Verbal Base Grammar' and Middle Korean verbal conjugation. *Morphology* 13.2: 245–265.

Kim, Yu Mi. 2002. A study of error analysis of Korean learners by using "Learner Corpus". *Teaching Korean as a Foreign Language* 27: 141–168.

Kim, You Jeong. 2005. Standards for error analysis of corpus by learners of Korean as a second language. *Journal of Korean Language Education* 16.1: 45–75.

Kim, Mikyung. 2017. An analysis of the Chinese KFL learners' errors and usage patterns of Korean particles. Master's thesis, Yonsei University.

Ko, Seok Ju, Sang Kyu Seo & Yun Jin Nam. 1999. Development of basic vocabulary semantic frequency dictionary for Korean language education. *Research on Language and Information* 1: 331–358.

Ko, Seok Ju, Mi Ok Kim, Je Yeol Kim, Sang Kyu Seo, Hui Jeong Chung & Song Hwa Han. 2004. *Korean learners' corpus and error analysis*. Seoul: Hankookmunhwasa.

Landis, J.R. & G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33.1: 159–174. https://doi.org/10.2307/2529310

Lee, Jeong Hui. 2002. A study on error determination standard and classification in Korean education. *Journal of Korean Language Education* 13.1: 175–197.

Lee, Yong. 2013. An efficient teaching method for Korean irregular verbs by analogy. *Teaching Korean as a Foreign Language* 38: 163–195.

Lee, Sun-Hee, Markus Dickinson & Ross Israel. 2013. Corpus-based error analysis of Korean particles. In *Proceedings of Learner Corpus Research*. Corpora and Language in Use Series 1. Louvain University Press. Belgium.

Lee, Jin Ho. 2014. *Lecture on Korean phonology*. Seoul: Samkyungmunhwasa.

Lee, Keonhui. 2019. A study on irregular errors analysis of predicates focused on corpus of Korean beginning learners. *Bilingual Research* 76: 57–81.

Lee, Chanyoung. 2018. Description of verb conjugation based on paradigmatic relations and 'regularity'. *Morphology* 20.1: 29–78.

Lee, Chanyoung. 2020. A study on error analysis of verb conjugation of Korean beginning learners. *Language and Culture* 16.2: 109–134. https://doi.org/10.18842/klaces.2020.16.2.5

Lennon, Paul. 1991. Error: Some problems of definition, identification, and distinction. *Applied Linguitiscs* 12: 180–196. https://doi.org/10.1093/applin/12.2.180

Lim, Seok-kyu. 2004. Restructuring via reanalysis in verbal paradigms. *Morphology* 6.1: 1–23.

Liu, Wenming. 2019. Error analysis of objective case postpositions of Chinese Korean learners. *The Language and Culture* 15.1: 223–249. https://doi.org/10.18842/klaces.2019.15.1.9

National Institute of Korean Language. 2018. *2018 Project on research and construction of the Korean Learner Corpus* (2018-01-49). Retrieved from https://korean.go.kr/front/reportData/reportDataView.do?mn_id=207&report_seq=956&pageIndex=1

Park, Mi Young, So Young Lee & Hyun Jin Lee. 1999. A study of the development of grammar test items on the basis of student's error. *Journal of Korean Language Education* 10.1: 141–171.

Selinker, Larry. 1972. Interlanguage. *Product Information International Review of Applied Linguistics in Language Teaching* 10: 209–241. https://doi.org/10.1515/iral.1972.10.1-4.209

Seo, Hyuk. 1992. Errors of learners of Korean and direction of textbook composition. *Senchengemwun* 20.1: 265–292.

Seo, Sang Kyu, Hyun Kyung Yu & Yun Jin Nam. 2002. Korean learner's corpus and Korean education. *Journal of Korean Language Education* 13.1: 127–156.

Shannon, Claude. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, Claude. 1951. Prediction and entropy of printed English, *The Bell System Technical Journal* 30: 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Stump, Gregory T. & Raphael A. Finkel. 2013. *Morphological typology: From word to paradigm.* New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139248860

Wang, Hye Sook. 2003. Korean language education using learner error corpus: United States, In Proceedings of 3rd Korean language education workshop.

Woo, Iin Hye. 1997. A study on the error of learners of Korean. *The Society of Korean Language & Culture* 15: 145–170.

Zhang, Jiemin. 2018. A study of displacement errors on Korean postposition *ey* and *eyse*. Master's thesis, Yonsei University.

Zhang, Jinxi & Hyounwha Kang. 2018. An analysis of adverbial particle errors in the Chinese Korean learners' corpus. *The Language and Culture* 14.3: 177–202. https://doi.org/10.18842/klaces.2018.14.3.8

Appendix A. Predictive entropy for all verb conjugation class in Korean

Rank	Conjugation class	Lexeme	Conjugation paradigm	Predictive entropy
1	ㄹ-class	SALTA	*salko, sani, sala*	1,049.30
2	ㅂ-class3-A	POYPTA	*poypkko, poyni, poye*	1,032.88
3	ㅡ-class	KKUTA	*kkuko, kkuni, kke*	412.43
4	ㅏ-class1	KATA	*kako, kani, ka*	313.73

Rank	Conjugation class	Lexeme	Conjugation paradigm	Predictive entropy
5	ㅜ-class2	PHWUTA	*phwuko, phwuni, phe*	306.29
6	ㅓ-class1	SETA	*seko, seni, se*	288.72
7	ㅎ-class1	NAHTA	*nakho, nauni, naa*	116.05
8	ㅎ-class2	NOLAHTA	*nolakho, nolani, nolay*	108.27
9	ㅅ-class2	NASTA	*natkko, nauni, naa*	102.83
10	ㅣ-class-C	CHITA	*chiko, chini, chye*	100.63
11	ㅂ-class2	TOPTA	*topkko, towuni, towa*	100.25
12	ㄿ-class	ULPHTA	*upkko, ulphuni, ulphe*	97.91
13	ㅍ-class	KAPHTA	*kapkko, kaphuni, kapha*	96.06
14	ㅄ-class	EPSTA	*epkko, epssuni, epsse*	95.81
15	ㅂ-class1	CAPTA	*capkko, capuni, capa*	74.34
16	ㄷ-class2	TUTTA	*tutkko, tuluni, tule*	72.48
17	ㅣ-class-B	CHITA	*chiko, chini, chiye*	67.30
18	ㄼ-class2	PALPTA	*papkko, palpuni, palpa*	65.23
19	ㅓ-class2	KULETA	*kuleko, kuleni, kulay*	55.66
20	ㅅ-class1	PESTA	*petkko, pesuni, pese*	46.04
21	ㅆ-class	ISSTA	*itkko, issuni, isse*	45.65
22	ㅊ-class	CCOCHTA	*ccotkko, ccochuni, ccocha*	45.38
23	ㅈ-class	NACTA	*natkko, nacuni, naca*	45.34
24	ㅌ-class	KATHTA	*katkko, kathuni, katha*	45.29
25	ㄷ-class1	TATTA	*tatkko, tatuni, tata*	45.20
26	ㅐ-class-B	KAYTA	*kayko, kayni, kay*	44.67
27	ㄴ-class	ANTA	*ankko, anuni, ana*	42.98
28	ㄼ-class1	NELPTA	*nelkko, nelpuni, nelpe*	39.00
29	ㄾ-class	HALTHTA	*halkko, halthuni, haltha*	39.00
30	ㄺ-class	MALKTA	*malkko, malkuni, malka*	39.00
31	르-class1	HULUTA	*huluko, huluni, hulle*	38.54
32	ㅂ-class3-B	POYPTA	*poypkko, poyni, pway*	38.16
33	ㄱ-class	NOKTA	*nokkko, nokuni, noka*	33.76
34	ㄲ-class	KYEKKTA	*kyekkko, kyekkuni, kyekke*	33.46
35	ㄵ-class	ANCTA	*ankko, ancuni, anca*	33.38
36	ㅁ-class	NAMTA	*namkko, namuni, nama*	33.33
37	ㄻ-class	KWULMTA	*kwumkko, kwulmuni, kwulme*	33.33

Rank	Conjugation class	Lexeme	Conjugation paradigm	Predictive entropy
38	ㅂ-class3-C	YECCWUPTA	*yeccwupkko*, *yeccwuni*, *yeccwe*	32.44
39	ㅏ-class2-B	HATA	*hako*, *hani*, *hay*	29.25
40	ㅀ-class	ILHTA	*ilkho*, *iluni*, *ile*	22.45
41	르-class2	PHWULUTA	*phwuluko*, *phwuluni*, *phwulule*	20.48
42	ㅕ-class	KHYETA	*khyeko*, *khyeni*, *khye*	19.03
43	ㅜ-class1-A	CWUTA	*cwuko*, *cwuni*, *cwue*	16.24
44	ㅚ-class-C	KOYTA	*koyko*, *koyni*, *kway*	15.65
45	ㅎ-class3	HAYAHTA	*hayakho*, *hayani*, *hayay*	14.06
46	ㅏ-class2-A	HATA	*hako*, *hani*, *haye*	13.19
47	ㅔ-class-B	PEYTA	*peyko*, *peyni*, *pey*	10.37
48	ㄶ-class	MANHTA	*mankho*, *manuni*, *mana*	8.86
49	ㅚ-class-B	KOYTA	*koyko*, *koyni*, *koyye*	7.32
50	ㅗ-class-A	POTA	*poko*, *poni*, *poa*	6.34
51	ㅚ-class-A	KOYTA	*koyko*, *koyni*, *koye*	5.41
52	ㅜ-class1-B	CWUTA	*cwuko*, *cwuni*, *cwe*	3.41
53	ㅣ-class-A	CHITA	*chiko*, *chini*, *chie*	2.35
54	ㅗ-class-B	POTA	*poko*, *poni*, *pwa*	2.15
55	ㅟ-class-B	CWITA	*cwiko*, *cwini*, *cwiye*	0.18
56	ㅞ-class-B	KKWEYTA	*kkweyko*, *kkweyni*, *kkwey*	0.18
57	ㅐ-class-A	KAYTA	*kayko*, *kayni*, *kaye*	0.00
58	ㅔ-class-A	PEYTA	*peyko*, *peyni*, *peye*	0.00
59	ㅞ-class-A	KKWEYTA	*kkweyko*, *kkweyni*, *kkweye*	0.00
60	ㅟ-class-A	CWITA	*cwiko*, *kwini*, *kwie*	0.00
61	ㅞ-class-C	KKWEYTA	*kkweyko*, *kkweyni*, *kkwey*	0.00
62	ㅟ-class-C	CWITA	*cwiko*, *cwini*, *cyuye*	0.00

Appendix B. Overall information on Korean Learner Corpus

Table 1. The main subjects of writing materials

Level	Subjects
1	Family, friends, hobbies, personality, likes and dislikes, and dreams
2	My family, my friends, and my neighbors
3	Memorable trip and travel experience
4	Recommended travel destination
5	Successful life for me
6	My thoughts about marriage

Table 2. The main subjects of spoken materials

Step	Subjects
Introduction	Self-introduction, small talk
Development	1. Hobbies, personality, likes and dislikes, family, and friends 2. Childhood, memorable episode in school years, hometown introduction, and recommended places 3. Dreams, ideas of success, thoughts about marriage, and what they want to do before they die

Table 3. Frequency and proportion by nationality

Nationality	Token frequency	Proportion (%)
China	129,297	22.79
Japan	104,754	18.47
Vietnam	91,866	16.20
United States	40,390	7.12
Taiwan	39,041	6.88
Russia	28,168	4.97
Others	133,710	23.57
Total	**567,226**	**100**

Note. "Others" included Thai (39,041 cases; 6.00%), Malaysia (12,329 cases; 2.17%), Kazakhstan (11,366 cases; 2.00%), and 71 more nations.

Table 4. Frequency and proportion by language

Nationality	Token frequency	Proportion (%)
Chinese	174,197	30.71
Japanese	106,248	18.73
Vietnamese	91,997	16.22
English	75,528	13.32
Russian	41,966	7.40
Thai	22,900	4.04
Others	54,390	9.59
Total	**567,226**	**100**

Note. "Others" included Spanish (5,478 cases; 0.97%), French (4,968 cases; 0.88%), Kazakh (4,527 cases; 0.80%), and 42 more languages.

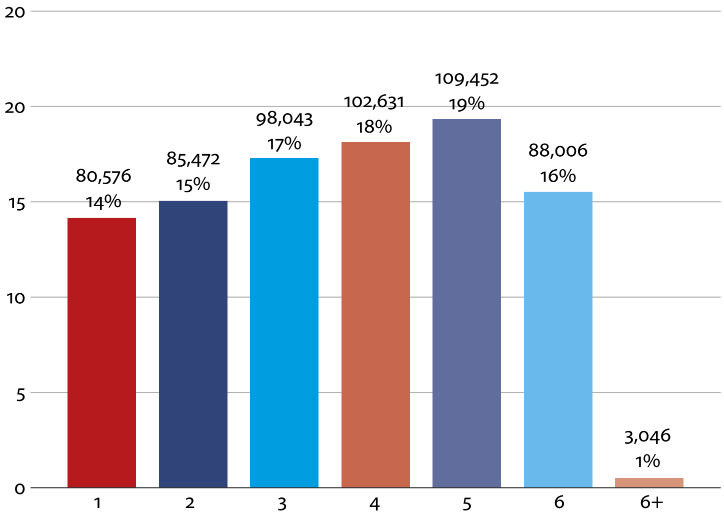


Figure 1. Frequency and proportion by language

Table 5. Frequency and proportion by location of errors

Location of error	Frequency	Proportion (%)
Common noun	18,594	18.62
Adverbial case particle	8,222	8.23
Verb	7,771	7.78
Connective ending	7,061	7.07
Subject case particle	6,782	6.79
Formulaic expressions	6,418	6.43
Others	45,014	6.00
Total	**99,862**	**100**

Note. "Others" included Object case particle (5,994 cases; 6.00%), Auxiliary particle (5,796 cases; 5.80%), Adnominal ending (3,986 cases; 3.99%), and 25 more locations.

Table 6. Frequency and proportion by pattern of errors

Pattern of error	Frequency	Proportion (%)
Substitution	39,940	49.99
Wrong form	19,445	24.34
Omission	14,059	17.60
Addition	6,456	8.08
Total	**79,900**	**100**

Address for correspondence

Chanyoung Lee
50, Yonsei-ro, Seodaemun-gu
Seoul
Republic of Korea
cy.lee@yonsei.ac.kr

Publication history

Date received: 15 March 2020
Date accepted: 5 December 2020