IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 October 2022, accepted 1 November 2022, date of publication 4 November 2022, date of current version 14 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3219448

RESEARCH ARTICLE

K-NCT: Korean Neural Grammatical Error Correction Gold-Standard Test Set Using Novel Error Type Classification Criteria

SEONMIN KOO[1], CHANJUN PARK[1,2], JAEHYUNG SEO[1], SEUNGJUN LEE[1], HYEONSEOK MOON[1], JUNGSEOB LEE[1], AND HEUISEOK LIM[1]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Upstage, Yongin, Gyeonggi-do 16942, Republic of Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT, South Korea, under the Information Technology Research Center Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation, under Grant IITP-2018-0-01405; and in part by the Basic Science Research Program through the National Research Foundation of Korea, Ministry of Education, under Grant NRF-2022R1A2C1007616.

ABSTRACT Recently, active research has been conducted on Korean grammatical error correction on machine translation (MT) and automatic noise generation. However, there is no gold-standard test set for objective and official comparative analysis. A significant limitation is measuring the ill-defined performance because the experimental error types in the train set are also included in the test set. Moreover, error types in the training set are also included in the test set. Additionally, the types of errors for qualitative analysis are defined differently with no explicit guidelines. This study proposes a gold-standard test set called the Korean Neural Grammatical Correction Test set (K-NCT) for Korean grammatical error correction using a new error type classification guideline. To ensure the factuality and reliability of the proposal, we conduct a quantitative analysis using a commercialization system and human evaluation. Experimental results demonstrate that the proposed grammatical error correction test set has a well-balanced, diverse, and precise guideline. Our dataset is available at https://github.com/seonminkoo/K-NCT

INDEX TERMS Korean grammar correction, error standard, gold test set, human evaluation.

I. INTRODUCTION

Grammatical error correction is a system that detects errors in a given sentence and corrects them. Particularly, in Korean, several grammatical errors occur owing to morphological richness and agglutinative characteristics [1], [2], [3].

The most intuitive solution for the Korean grammatical error correction is a rule-based approach. In this approach, several error types and their corresponding correcting rules are predefined for the correction process [4]. This method is effective because spelling and grammatical errors are amended without destroying the original sentence structure, hence this approach is currently utilized. However, this method has several limitations in that it is time consuming and requires human resources to establish the correction rules. Furthermore, it strictly revises sentences by the predefined rules and cannot correct other types of errors.

To address these problems, a statistic-based approach is proposed, which mitigates the necessity for the construction of correction rules that require high resources and judge errors based on the probability estimated by the given corpus [5], [6]. However, it also demands a sufficiently large corpus to attain decent performance.

Recently, deep learning-based grammatical correction algorithms that can effectively alleviate the above limitations are being utilized. Several methods that can construct an error correction model without parallel data have been proposed, particularly for the Korean language [1], [7], [8]. These are mostly based on the automatic noise injection process that generates pseudo-parallel corpus through the unlabelled

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

mono corpus [1]. Generally, a sentence given by mono corpus is regarded as a target sentence, and through the noising process, a corresponding source sentence is generated for the training of sequence to sequence-based correction model.

However, despite the significant improvement in the deep learning-based approach, these approaches contain the following two limitations. First is the absence of the official grammatical error correction dataset. This follows the inconsistent evaluation of the error correction model. Because public test data does not exist, researchers construct their test sets using the arbitrary sampling of their original corpus. As test data differs according to each research study, the corresponding performance assessment may not reflect the objective and reliable performance of each model [9], [10].

Second, the unstandardized error types used in each research study worsen the performance assessment's reliability. The high performance of each model may not reflect its effectiveness because a precise standard for the error types has not been established [2]; hence it may be underestimated or overestimated by the utilized test set.

We propose the error type classification standards for Korean grammatical error correction research and release the corresponding gold-standard test set, K-NCT (Korean Neural grammatical Correction Test set). The proposed types include four significant criteria: spacing, punctuation, numerical, spelling and grammatical error. These are divided into 23 subcategories related to balance, diversity, and factuality. We proceed with human evaluation by consulting linguistic experts and qualitative analysis through publicly released commercialization systems to secure the factuality and reliability of the K-NCT.

Section III.B describes the three reliability features (balance, diversity, and factuality) considered for K-NCT construction. Subsequently, we propose an error type criteria reflecting the characteristics of Korean in Section III.C. Section III.D describes the data selection processes, preprocessing, post-processing, and error injection to construct K-NCT. Finally, Section III.E presents an overview of the completed K-NCT. IV describes the experiments and results.

The contributions of this study are as follows.

- It identifies the limitations of the performance evaluation methods that are utilized in previous studies and suggests a ''New Error Type Classification Criteria.'' Presumably, it is the first error type standard.
- Based on the error type criteria, we released the first gold-standard test set for Korean grammatical error correction called K-NCT for active Korean grammatical error correction research.
- An in-depth quantitative analysis is performed by applying it to the commercialization system, and objective and reliable K-NCT verification is conducted through human evaluation.

II. RELATED WORKS

Several deep learning-based Grammatical error correction models solve the task from the point of view of machine translation. From a machine translation perspective, grammatical error correction is ''translating'' an error sentence into a correct sentence. It is mainly corrected through the noising encoder and denoising decoder structures based on the sequence-to-to-sequence model [11], [12], [13].

Recently, research on grammatical error correction for high-resource languages is active. A sequence tagging model including a transformer-based encoder is proposed for error detection and error correction [14]. From an educational point of view, interpretability is improved by adding text and examples for language learners expanding the reason for correction together [15]. Self-Supervised Curriculum Learning is applied to measure data difficulty through training loss and train the model to increase performance [16]. Applying a contrastive learning approach to the GEC model improves performance for low error density domains [17]. Considering the inference efficiency, decoding many tokens through aggressive decoding improves the model's speed [18].

However, to apply the methodology, a parallel corpus composed of pairs of error sentences and correct sentences is required; nonetheless, there is no publicly available Korean data. Various studies are underway to construct a pseudo parallel corpus without human-labor by applying the automatic noise generation technique. The automatic noise generation technique automatically generates a parallel corpus by designing a noise function for a mono corpus and applies it to generate a parallel corpus. Grapheme-to-phoneme, spacing, punctuation, and pronunciation errors, etc. are artificially injected to add noise. However, because the types of errors defined for each study are different, and the training set and test set are divided and used in the pseudo generated data set. The type of error used for training will be high probability included in the test, rendering objective research difficult.

Therefore, this study pointed out the limitations of data construction and performance evaluation of the existing Korean grammatical correction research and proposed the error type classification system for the first time. Also, based on the system, presumably, a gold-standard test set for Korean grammatical error correction system called K-NCT was built for the first time.

III. K-NCT

A. WHY K-NCT?

Gold-standard test set has not been used in previous studies for Korean grammatical error correction. This causes several limitations.

First, it is impossible for accurate performance measurement to generate pseudo-parallel corpus, which includes only certain types. Second, the error type used in training is included in the test because the generated corpus is divided arbitrarily. Third, it is difficult to analyze model development in detail because there is no systematic error type guideline. Fourth, some datasets use single language and sentences of domains and lengths that are not diverse; therefore, objective comparative studies are complex.

We propose K-NCT set, which was constructed by considering the error type classification criteria and various reliable factors to solve the problems above. K-NCT set not only systematizes error types and applies these to actual sentences; it also considers different domains, methodologies, and the number of syllables. It is a 100% human-constructed high-quality dataset.

B. DESIGNING K-NCT

K-NCT can be applied to various features of constructs that can be objectively verified. First, the dataset is designed considering the balance, which produces fair features and comprises data that are unbiased. Second, it is an organized dataset that considers diversity; therefore, the dataset assumes numerous features. Third, the dataset is created by reflecting the factuality. The objective gold-standard test set has little to no unreal data and unnatural error types for evaluating the model by humans.

a: CONSIDERING THE BALANCE

K-NCT includes well-balanced error types, syllable lengths, text style, and domains. The balance is achieved by determining the proportion of error types and ratios based on details in the case of spelling and grammatical errors. It has a fairness configuration of 500 spacings, 500 punctuations, and 500 numerical errors. The detailed error types of spelling and grammatical errors consist of 1312 monolingual, 200 multilingual, 800 spelling, 411 syntax, 200 semantic, and 100 neologism errors. This balance is adjusted by dividing the length of the syllable into a specific range and setting the ratio for each degree. Text styles compose sentences of various types considering the characteristics of Korean and proportions of each to be uniform.

b: CONSIDERING THE DIVERSITY

We consider not only single text style, but also written, spoken, and dialog style for the diversity of K-NCT. Several previous datasets and pseudo-parallel corpus only contain single text style in certain previous studies. Therefore, we include various text styles to construct the dataset for a more accurate and objective comparative analysis.

The dataset covers the range of syllable lengths of 2∼20, 21∼29, 30∼50, and 51 syllables or more to ensure diversity. Furthermore, it proposes various types of errors and constructs the dataset that satisfies them to secure the errors. It is possible to determine which error types are bullish and bearish in a model through multiple error types.

c: CONSIDERING THE FACTUALITY

K-NCT conducts a human evaluation to prove the factuality of the proposed guidelines and generated sentences (see Section 3.3). Because 'factuality' is a naturalist or true to life for a person to judge, it undergoes a human evaluation. We present human evaluation criteria to determine the factuality and corresponding score.

C. ERROR TYPE CLASSIFICATION CRITERIA FOR K-NCT

Accurate performance evaluation and analysis of Korean grammatical error correction models requires diverse and systematic error types. To this end, a more detailed and specific error type system is proposed. Table 1 shows the error type classification criteria used in K-NCT construction. It is classified into 23 detailed error types based on four major categories (spacing, punctuation, numerical, and spelling and grammatical errors). Particularly, spelling and grammatical error is divided into primary and secondary errors for detailed analysis.

a: SPACING ERROR

This violates the Korean spacing rules. Similar to deep learning models, humans also commit these errors because of their fast typing speed or habits.

b: PUNCTUATION ERROR

This occurs when punctuation marks are not attached in the sentence or misplaced. When generating a sentence with deep learning model, it easily appears because of an unregistered word. The intent of the sentence may be different depending on punctuation marks.

c: NUMERICAL ERROR

This occurs when a cardinal number indicates quantity and an ordinal number indicates order. For example, the correct sentence '한 시 일 분(han si il bun)' is incorrectly written as '하나 시 일 분(hana si il bun)' or '일시 일분(ilsi ilbun)'.

d: SPELLING AND GRAMMATICAL ERROR

This violates Korean spelling and grammatical. As the most frequent case in Korean, it is divided into primary and secondary errors. Primary and secondary errors can be nested. Therefore, the primary error is first classified and the second error is sub-classified.

Primary errors are classified into monolingual errors occurring in Korean and multilingual errors occurring in other languages. The subtypes are as follows:

- **Remove error**: This occurs when some words are not recognized, or the endings or postpositions are omitted. This is one of the common mistakes Koreans commit, and it is classified as an error type.
- **Addition Error**: This occurs when the same word is repeated, postpositions are not used, or when endings are added. These mistakes are committed frequently owing to sentence typing speed and incorrect grammatical knowledge.
- **Replace Error**: It is subdivided into word replacement, in which another word replaces a word, and rotation replacement, in which the order of syllable changes within one phrase. The errors occur at a fast typing speed, primarily related to spelling errors around the intended spelling position.

TABLE 1. Proposed novel error type classification criteria for Korean grammatical and spelling error correction.

Error Type					Explanation
Spacing Error					Violating the spacing rules
Punctuation Error					Punctuation marks are not attached in Korean sentences or are attached in the wrong position
Numerical Error					Cardinal number indicating quantity and the ordinal number indicating the order are in error
Spelling and Grammatical Error	Primary Error	Monolingual Error	Remove Error		Some words are not recognized, or endings or suffixes are omitted
			Addition Error		Same word is repeated, or an unused postposition or ending is added
			Replace Error	Word replace	Word is replaced by another word
				Rotation replace	Order of syllables changes within a one phrase
			Separation Error		Separating consonants and vowels in characters
		Multilingual Error	Typing language Error		Typing while the keyboard is not in Korean mode
			Foreign word conversion Error		Writing differently from the standard foreign language pronunciation
	Secondary Error	Spelling Error	Consonant vowel conversion error		Spelling error in non-speaking alphabet units
			Grapheme-to-phoneme(G2P) Error		Writing spellings according to pronunciation
		Syntax Error	Element Error		The Korean sentence components are not equipped or the word order is not correct
			Tense Error		Using a verb that does not match the tense
			Postposition Error		Probing that does not fit the grammar
			Suffix Error		Using an ending that is not grammatically correct
			Auxiliary predicate Error		Using an auxiliary verb that is not grammatically correct
		Semantic Error	Dialect Error		Writing in non-standard language
			Polite speech Error		An adjective expression that does not fit the subject
			Behavioral Error		Expressions that the subject cannot perform
			Coreference Error		Invalid entity reference
			Discourse context Error		Contradicting the context of the previous discourse
		Neologism Error			Using grammar or new words that are not included in the existing grammar system

- **Separation Error**: This occurs when consonants and vowels of a character are separated. This frequently occurs in Korean because the space key is usually used to separate words with spaces.
- **Typing language Error**: This occurs when typing while the keyboard is not in Korean mode. It occurs when the language change key is pressed while manipulating the keyboard, for example, '안녕' is incorrectly typed as 'dkssud'.
- **Foreign word conversion Error**: This generates differently from the standard foreign language pronunciation. In Korean, there is a normal foreign language notation; however, it is the case that does not follow it, for example, '수프(supeu)' is incorrectly spelled '스프(seupeu).'

Secondary error is classified into four types (Spelling, Syntax, Semantic, and Neologism Errors). The subtypes are as follows:

- **Consonant vowel conversion error**: Spelling error in non-speaking alphabet units, for example, '이제 곧 갑니다.(ije god gabnida.)' is incorrectly written as '이제콘 갑니다.(ije kon gabnida.)'
- **Grapheme-to-phoneme(G2P) Error**: Writing spellings according to pronunciation, for example, '이제콘 갑니다.(ije god gabnida.)' is incorrectly written as '이제 곧 갑니다.(ije goj gabnida.)'
- **Element Error**: Korean sentence elements are not in place or do not fit in the order of words. Korean has a fixed sentence structure (subject, object, verb), and there are transitive verbs that require an object.
- **Tense Error**: Using a verb that does not match the tense, e.g., 'Did in the future.'.
- **Postposition Error**: Using a postposition that does not conform to the grammatical. Because Korean is an agglutinative language, the use of the verb is important. Furthermore, there are various types of verbs such as case and auxiliary verbs.
- **Suffix Error**: Using endings that do not conform to spelling. Suffix Error is classified as an error type because the original verb is modified according to each situation.
- **Auxiliary predicate Error**: Using auxiliary verbs that do not conform to grammar. This occurs because Korean uses auxiliary verbs to construct sentences.
- **Dialect Error**: Writing in non-standard language. Korean has a variety of dialects. The criterion for determining dialect errors is the speaker or author's intention. When the model fails to create the intended dialect, it is judged as an error.
- **Polite speech Error**: Polite speech expression that does not fit the subject. This error type reflects Korean cultural characteristics.
- **Behavioral Error**: An expression that the subject cannot perform, for example, 'the apple eats the banana'.
- **Coreference Error**: Invalid entity reference. Misrecognition of an object can create a sentence different from the intended one.
- **Discourse context Error**: Contradicting the context of the previous discourse. This error encourages the generation of a wrong sentence with different information from the previous sentences.
- **Neologism Error**: Using a spelling or a new word that is not in the existing grammar system. Similar to dialect

TABLE 2. Example of K-NCT test data sample.

K-NCT Test Data Sample
{ "index": 408, "error sentence": "지금 내가 그녀를 찾고 있는데 <e1>보이지</e1><e2>아니아.</e2>", "correct sentence": "지금 내가 그녀를 찾고 있는데 보이질 않아.", "domain": "daliy", "style": "spoken", "syllable": 18, "phrase": 7, "number of error": 2, "error type":{"e1": "word_replace,suffix", "e2": "word_replace,auxiliary_predicate"}}

error, neologism error is the speaker's or author's intention used as a criterion for judging errors.

By establishing this guideline, we built K-NCT, a test set consisting of sentences that reflect a detailed error type classification system. This allows accurate performance measurements and clear comparative studies. In this study, we applied K-NCT to actual Korean sentences to generate sentences that included various error types., We conducted an experiment based on the guidelines to evaluate the performance and verify the reliability of K-NCT as a gold test set.

D. CONSTRUCTION PROCESS

a: DATA SELECTION

All original sentences are extracted from the Korean-English translation (parallel) corpus of AI hub[1] [19]. AI hub is a data platform operated nationally in Korea. It processes natural language datasets such as machine translation or document summarization, as well as various fields such as images, autonomous driving, and healthcare datasets. In summary, AI hub builds high-quality and large-capacity datasets in the Korean language and contributes to AI research by disclosing those datasets to the public.

In the case of Korean-English translation (parallel) corpus data, including those files that are written, spoken, and dialog styles, the written style is a formal language mainly used in formal occasions, and the spoken and dialog styles are relatively informal and natural languages. The dataset consists of various domains and syllables.

Written, dialog, and spoken styles consist of 800,000, 100,000, and 400,000 raw sentences, respectively.

We integrate the individual documents by style to create three files consisting of 1,100,000 written style sentences, 400,000 spoken style sentences, and 100,000 dialog sentences. Considering balance and diversity, we randomly sampled 1,000 sentences from each of the three language style files and collected the dataset as 3,000 sentences. This dataset uses as raw data for the gold-standard test set.

b: PRE-PROCESSING

Pre-processing is the first step to conduct error injection in correct sentences. Accordingly, we performed sentence alignment and error tagging.

First, the sentence alignment is performed, that is, aligning sentences considering the reliability of syllables. The sentences are sorted in the corresponding syllable range based on the index. Index 0∼749 are arranged to correspond to 2∼20 syllables, 750∼1799 correspond to 21∼29 syllables, 1800∼2399 correspond to 30∼50 syllables, and 2400∼2999 correspond to more than 51 syllable. Because the average number of syllables in the dataset is different for each domain, it is difficult to control the number of syllables in the downstream domain, which hinders the reliability. For factuality, we do not consider the domain when we sort the sentences by the range of syllables.

Second, error type tagging is performed, that is, tagging each sentence with an error type. We randomly tagged the spacing, punctuation, numerical, monolingual, multilingual, spelling, syntax, semantic, and neologism errors on the sentence of the dataset by a predefined ratio.

Some of the error types require essential conditions to occur. For example, a numerical expression must include a numerical error in the sentence. Also, dialect or neologism errors must include a dialect or neologism error in the sentence. Therefore, if the essential condition of the tagged error type is not satisfied, we switch the alignment position of the error type with another sentence that satisfies the condition. Although this method is quite simple, it required post-processing because it cannot be applied to all error types.

c: POST-PROCESSING

Post-processing is performed after pre-processing, in which correct sentences that do not have fitted error conditions are targeted. We proceeded with the correct sensation modification process to meet the error type generation conditions.

In the switching method of pre-processing, there are sentences that have no conditions for generating errors. Therefore, we form high-quality data by modifying the correct sentences such that the essential conditions for error type insertion are satisfied while keeping Korean orthography. For example, there are cases where extracted original sentences do not contain dialects and neologisms. In such cases, we corrected the sentences to include dialect or neologism. In addition, if the numerical expressions do not match the proposed format, they are corrected. Using post-processing, we generated high-quality data that can satisfy these conditions.

d: ERROR INJECTION

The error types included in the correct sentence were sorted through three steps (data selection, pre-processing, post-processing). Error Injection generates an error sentence (incorrect sentence) including the sorted error type in the correct sentences. The labor generates incorrect sentences by performing error injection according to the given correct sentence and error type. The guideline for error injection is as follows.

- Define the error type and present actual examples.
- Modify the number of syllables in index, giving limited freedom.
- Restrict changes in the style of the sentence.

[1] https://bit.ly/3QwW9IT

- Number expression represent only statistical or date expression, other than in letters.
- When correcting spelling, use the part within two editing distance of the keyboard.
- Indicate span and the corresponding error type when the error occurs.

Labor must have a certain understanding of the error types to generate high-quality data. Therefore, definitions of error types and practical examples are given to labor. Labor generates error sentences within the range of syllables of the specified index, forcing the text-styles to stay constant because they were pre-processed before providing them to the generator.

The guideline presents a numerical expression that can generate a numerical error. Statistical or date expression may be generated as a number, and in other cases expressed as a letter. For example, 'November 2021' is presented as a number and 'there are two apples' is presented as a character. The candidates for the changed character are limited to the characters whose keyboard editing distance is two or less to reflect the factuality of the error sentence.

Error types are randomly tagged at sentences by designating a balanced number for each error type. The labor identifies the correct sentence for the corresponding error type and creates an error sentence. Through the guideline presented above, a certain standard of freedom allows the creation of more realistic dataset. In addition, labor directly inspects all data for the accuracy and reliability of the K-NCT.

E. FINAL GOLD TEST SET

Three thousand high-quality sentences are constructed through data selection and two pre-processing (error type alignment), post-processing, and error injection processes. Six people[2] created and inspected the guidelines and built and evaluated the data to build the dataset. Each dataset is specialized by completing training on detailed descriptions of the guidelines for more than three hours to create high-quality data. K-NCT constructs a json format file that is publicly released as the corpus.

Table 2 shows an actual example of the generated final gold-standard test set. K-NCT contains errors and correct sentences, applicable domains, phrases, a number of syllables, and a number of errors and error types. Error sentences mark a span at the location of the error, and the error type is indicated with the error type at the location. Because errors may occur in multiple locations in one sentence, the span may be displayed in the phrase where the error occurred, as shown in '⟨e1⟩ error ⟨/e1⟩', to display several error locations and types, as shown in the second row of Table 2.

IV. EXPERIMENTS AND RESULTS

We conducted experiments and validation to prove the reliability and objectivity of K-NCT. Three aspects (statistical, quantitative, and human evaluation) were utilized to verify the quality of K-NCT as a gold-standard test set.

[2]The anotators consist of ordinary people with no background knowledge of the task and are trained before starting the assessment.

TABLE 3. Statistics of our K-NCT dataset. # of sents/tokens/words: number of sentences/tokens/words; △ avg of SL/WS/SS: average of sentence length/words/spaces per sentence; ∗ # of K-toks/E-toks/S-toks: number of Korean/English/special-symbol letter tokens.

K-NCT	Test	
	Error sentence	Correct sentence
# of sents	3,000	3,000
# of tokens	129,798	129,886
# of words	31,183	31,700
avg of SL △	43.27	43.29
avg of WS	10.39	10.57
avg of SS	9.39	9.57
# of K-toks ∗	93,794	93,856
# of E-toks	904	860
# of S-toks	4,785	4,423

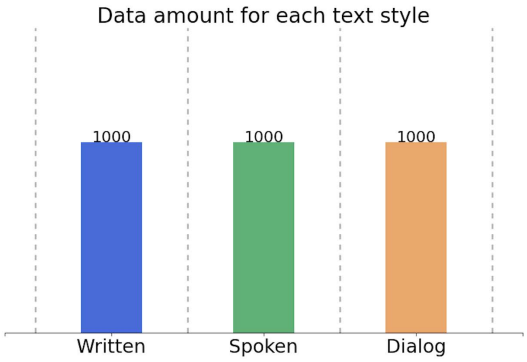

FIGURE 1. Data amount for each text style.

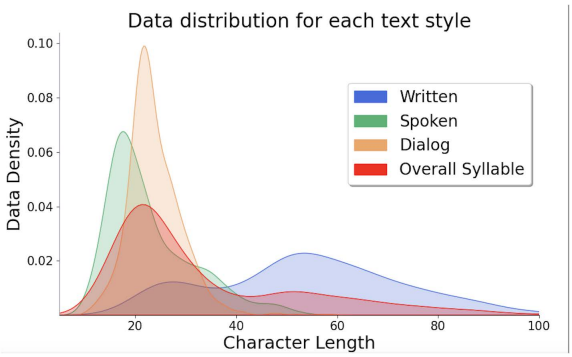

FIGURE 2. Data distribution for each style.

A. STATISTICAL ANALYSIS

First, we conducted basic statistical analysis as shown in Table 3. The K-NCT consists of 3000 sentences. We define the error sentences as source and the correct sentences as target. The length of error sentences on average is 43.27, the number of words on average is 10.39, and the number of spaces on average is 9.39. The length of correct sentences on average is 43.29, the number of words on average is 10.57, and the number of spaces on average is 9.57. Considering the statistical similarity between the error sentence and the correct sentence, these results show that the generated error

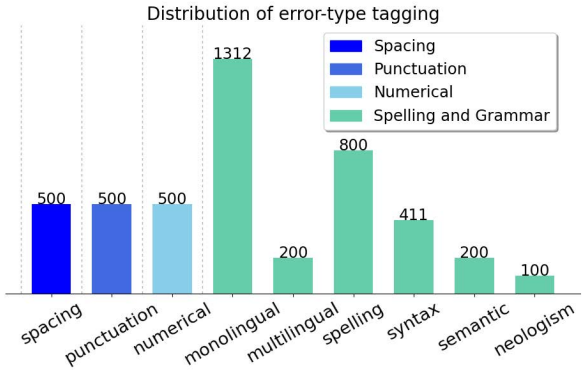

FIGURE 3. Distribution of error-type tagging on a sentence-by-sentence basis.

sentences consist of realistic errors that are sufficient to understand the original sentence.

Error sentences contain 93,794 K-tokens, 904 E-tokens and 4,785 S-tokens. Correct sentences contain 93,856 K-tokens, 860 E-tokens and 4423 S-tokens. In the real world, English is mixed with the Korean sentences. Through the punctuation error injection, we generated error sentences covering various punctuation.

Secondly, we analyzed the reliability of K-NCT. Figure 1 and 2 show distribution for each text style. Figure 1 shows the data amount for text style(written, spoken, dialog) that reflect Korean characteristics. Each of them consists of 1000 sentences, indicating that K-NCT satisfies the equality of style. Figure 2 shows the syllable distribution for each style. In the case of written style, 50∼70 syllables, spoken style 15∼25 syllables, and in dialog style, 20∼30 syllables are distributed.

This shows that K-NCT is composed of realistic sentences. All sentences of K-NCT are distributed in 15∼30 syllables, including other syllables. This depends on the situation and style in which the vocabulary is used, and the length of the syllable varies. Therefore, K-NCT is constructed considering the diversity of style.

Figure 3 shows the distribution of error type tagging. Considering the balance and diversity, 1500 sentences are tagged with spelling and grammatical errors for detailed error types, and the remaining types are tagged in 500 sentences each.

Monolingual error is tagged in 1312 sentences, and spelling error is tagged in 800 sentences. Multilingual and semantic errors are tagged in 200 sentences, syntax error is tagged in 411 sentences, and neologism error is tagged in 100 sentences. This shows that the various error types are well-balanced. Therefore, exploiting K-NCT can generate detailed measurements through high-quality sentences generated by considering reliability features (e.g., style, syllable, and error type).

B. QUANTITATIVE ANALYSIS

We conducted quantitative analysis. Based on K-NCT, we conducted a proofreading performance comparison

TABLE 4. Various evaluation scores on Korean grammatical corrections.

	GLEU	BLEU
Naver	69.83	70.29
Daum	71.96	73.63
Pusan	73.60	74.43

experiment for Naver,[3] Daum,[4] and Pusan,[5] which are the most representative models of the Korean grammatical correction commercialization system. The reason for choosing the commercialization system for comparison is that it is a certified system used by several researchers, and the latest deep learning-based grammatical correction methodology is applied; hence, it is the most objective and reliable system for accurate analysis.

The performance of each corrector is measured by using the error sentences of K-NCT as input for three commercialization systems and performing quantitative analysis using the BLEU score [20] and GLEU score [21], which are used in various deep learning-based grammatical correction studies as evaluation indicators [1], [22], [23]. The experimental results are shown in Table 4.

Experimental results show significant performance in the order of Pusan, Daum, and Naver based on the GLEU and BLEU scores. However, the performance difference is not significant for each subtype in a fairly similar section. K-NCT is a gold-standard test set that can objectively measure performance without biasing any system.

As an additional experiment, the strengths and weaknesses of each commercialization system are analyzed based on the error type classification criteria designed in this study. The experimental results are displayed in Table 5.

We clearly analyze each commercialization system used K-NCT. First, Pusan, which shows the best overall performance, shows an overwhelmingly better performance than other models in spacing. Pusan model shows high performance of 87.13 based on BLUE score and 86.07 based on GLUE score, whereas Naver and Daum show a significantly lower performance with a BLUE score of 60.9 and 56.30, and GLUE score of 44.02 and 40.83, respectively. It is found that the Pusan model is the most robust model for correcting spacing.

Second, in the case of punctuation, Daum shows the best performance with BLEU and GLUE scores of 69.16 and 50.61, respectively. In the case of numeric, daum based on BLEU score and Pusan based on GLEU score shows the best performance; however, there is no significant difference in all three models. Finally, in the case of spelling and grammatical, which is the most important performance, it is found that the Pusan model shows the best performance with BLEU score of 75.07 and GLEU score of 70.66.

As shown in the analysis above, because the error type for each sentence pair is labeled in K-NCT, the strengths and

[3]https://bit.ly/3CtaPl4
[4]https://alldic.daum.net/grammatical_checker.do
[5]http://speller.cs.pusan.ac.kr/

TABLE 5. Evaluation score of Korean grammatical error corrector commercialization system by error type.

	Naver				Daum				Pusan			
	Spacing	Numerical	Puncutation	Spell and Grammatical	Spacing	Numerical	Puncutation	Spell and Grammatical	Spacing	Numerical	Puncutation	Spell and Grammatical
GLEU	60.90	67.80	66.08	72.79	56.30	69.76	69.16	74.79	87.13	68.65	66.47	75.07
BLEU	44.02	53.16	47.28	58.48	40.83	54.77	50.61	58.8	86.07	58.46	48.13	70.66

TABLE 6. Score for each evaluation item in human evaluation. For human evaluation, we denote the average score for each quality indicator estimated by the five language experts. # of poor/excellent indicate ratio of poor/excellent, respectively.

	Human Evaluation	# of poors	# of excellent
Compositionality	1.53	0.03	0.53
Association	1.15	0.18	0.33
Fluency	1.50	0.05	0.55
Factuality	1.28	0.09	0.38
Typicality	1.30	0.15	0.48

weaknesses of the corresponding grammatical corrector can be clearly analyzed.

C. HUMAN EVALUATION

In order to evaluate the diversity and effectiveness of the generated dataset, a total of 200 sentences are randomly sampled for 10 error types, excluding Secondary Spelling and Grammatical Errors. Since Secondary Spelling and Grammatical Errors are a subset of Primary Errors, we do not sample them separately.

We employ 5 human evaluators with bachelor's degrees. Each evaluator performed the evaluation after receiving education on the introduction and evaluation method of K-NCT. The quality is evaluated for five items, and a score of zero for poor, one for normal, and two for excellent is given. The evaluation items are as follows [24]. (1) It contains all given error types (compositionality). (2) The relationship/use between error types is natural (Association). (3) It is easy to grasp the original meaning/intention of the sentences (Fluency). (4) It is easy to recognize which type of error the error sentence contains (Factuality). (5) It is a common error type (Typicality). The experimental results are as shown in Table 6.

As a result of the evaluation, the average score for the five evaluation items is high, with an average score of one point in the mid-range. Particularly, it is noteworthy that most of the ratio of excellent scored high. Through human evaluation, it did not only prove the realism of the error type; it also demonstrated that it consists of high-quality data.

V. CONCLUSION

Recently, many studies on grammatical error correction based on machine translation and automatic noise generation have been conducted. However, in the case of Korean grammatical error correction, an objective comparative study is difficult because a pseudo corpus including only specific error types are generated and used owing to the absence of a learning dataset and a gold test set. To solve this problem, new error type classification criteria are proposed, based on which K-NCT, a gold test set for Korean grammatical correction research, is built for the first time. In addition, the reliability of the proposed K-NCT is verified through statistical analysis, quantitative analysis, and human evaluation, and the data is completely accessible to the public. Our dataset can be applied not only to formal situations such as news articles, but also to dialogue or spoken Korean spelling correction tasks.

Our limitation is that Coreference and Discourse context Errors cannot be determined in sentence unit errors. Therefore, these error types are not included in the test dataset. In the future, to deal with errors that occur in paragraph units, the data will be expanded in units of paragraphs and to various language pairs. Based on documentation or conversation resources, we will construct and publish datasets including types of errors not included in this version of the dataset.

ACKNOWLEDGMENT

This is an expanded paper of "Classification and Analysis of Error Types for Deep Learning-Based Korean Spelling Correction" [DOI:10.15207/JKCS.2021.12.12.065]. *(Seonmin Koo and Chanjun Park contributed equally to this work.)*

REFERENCES

[1] C. Park, K. Kim, Y. Yang, M. Kang, and H. Lim, "Neural spelling correction: Translating incorrect sentences to correct sentences for multimedia," *Multimedia Tools Appl.*, vol. 80, pp. 1–18, Jun. 2020.
[2] Y. Wang, Y. Wang, J. Liu, and Z. Liu, "A comprehensive survey of grammar error correction," 2020, *arXiv:2005.06600*.
[3] J.-H. Lee and H.-C. Kwon, "Context-sensitive spelling error correction techniques in Korean documents using generative adversarial network," *J. Korea Multimedia Soc.*, vol. 24, no. 10, pp. 1391–1402, 2021.
[4] J. Xiong, Q. Zhang, S. Zhang, J. Hou, and X. Cheng, "Hanspeller: A unified framework for Chinese spelling correction," *Int. J. Comput. Linguistics Chin. Lang. Process.*, vol. 20, no. 1, pp. 1–22, Jun. 2015.
[5] M. Kim, J. Jin, H.-C. Kwon, and A. Yoon, "Statistical context-sensitive spelling correction using typing error rate," in *Proc. IEEE 16th Int. Conf. Comput. Sci. Eng.*, Dec. 2013, pp. 1242–1246.
[6] J.-H. Lee, M. Kim, and H.-C. Kwon, "Improved statistical language model for context-sensitive spelling error candidates," *J. Korea Multimedia Soc.*, vol. 20, no. 2, pp. 371–381, Feb. 2017.
[7] M. Lee, H. Shin, D. Lee, and S.-P. Choi, "Korean grammatical error correction based on transformer with copying mechanisms and grammatical noise implantation methods," *Sensors*, vol. 21, no. 8, p. 2658, Apr. 2021.
[8] C. Park, S. Park, and H. Lim, "Self-supervised Korean spelling correction via denoising transformer," in *Proc. 7th Int. Conf. Inf., Syst. Converg. Appl.*, 2020.
[9] A. Rozovskaya and D. Roth, "Grammar error correction in morphologically rich languages: The case of Russian," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 1–17, Mar. 2019.
[10] K. Imamura, K. Saito, K. Sadamitsu, and H. Nishikawa, "Grammar error correction using pseudo-error sentences and domain adaptation," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 388–392.
[11] H. Li, Y. Wang, X. Liu, Z. Sheng, and S. Wei, "Spelling error correction using a nested RNN model and pseudo training data," 2018, *arXiv:1811.00238*.
[12] A. Solyman, Z. Wang, and Q. Tao, "Proposed model for Arabic grammar error correction based on convolutional neural network," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Sep. 2019, pp. 1–6.
[13] A. Kuznetsov and H. Urdiales, "Spelling correction with denoising transformer," 2021, *arXiv:2105.05977*.
[14] M. Tarnavskyi, A. Chernodub, and K. Omelianchuk, "Ensembling and knowledge distilling of large sequence taggers for grammatical error correction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3842–3852. [Online]. Available: https://aclanthology.org/2022.acl-long.266

[15] M. Kaneko, S. Takase, A. Niwa, and N. Okazaki, "Interpretability for language learners using example-based grammatical error correction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7176–7187. [Online]. Available: https://aclanthology.org/2022.acl-long.496

[16] Z. Gan, H. Xu, and H. Zan, "Self-supervised curriculum learning for spelling error correction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3487–3494.

[17] H. Cao, W. Yang, and H. T. Ng, "Grammatical error correction with contrastive learning in low error density domains," in *Findings of the Association for Computational Linguistics: EMNLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4867–4874. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.419

[18] X. Sun, T. Ge, F. Wei, and H. Wang, "Instantaneous grammatical error correction with shallow aggressive decoding," 2021, *arXiv:2106.04970*.

[19] C. Park, M. Shim, S. Eo, S. Lee, J. Seo, H. Moon, and H. Lim, "Empirical analysis of Korean public AI hub parallel corpora and in-depth analysis using LIWC," 2021, *arXiv:2110.15023*.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, May 2001, pp. 311–318.

[21] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2015, pp. 588–593.

[22] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H. Lim, "BTS: Back transcription for speech-to-text post-processor using text-to-speech-to-text," in *Proc. 8th Workshop Asian Transl. (WAT)*, 2021, pp. 106–116.

[23] C. Park, Y. Yang, C. Lee, and H. Lim, "Comparison of the evaluation metrics for neural grammatical error correction with overcorrection," *IEEE Access*, vol. 8, pp. 106264–106272, 2020.

[24] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," 2020, *arXiv:2006.14799*.

[25] S. Koo, C. Park, A. So, and H. Lim, "Classification and analysis of error types for deep learning-based Korean spelling correction," *J. Korea Converg. Soc.*, vol. 12, no. 12, pp. 65–74, 2021.

SEONMIN KOO received the B.S. degree from the Department of Computer Science and Engineering, Konkuk University, Seoul, South Korea, in 2022. She is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, and the integrated master's and Ph.D. degrees. She is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. Her research interests include machine translation and knowledge base population.

CHANJUN PARK received the B.S. degree in natural language processing and creative convergence from the Busan University of Foreign Studies, Busan, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Korea University, Seoul, South Korea. From 2018 to 2019, he worked at SYSTRAN as a Research Engineer. He is also working as an AI Research Engineer at Upstage. His research interests include machine translation, grammar error correction, simultaneous speech translation, and deep learning.

JAEHYUNG SEO received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2020. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, and the integrated master's and Ph.D. degrees. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include language generation and decoding strategy, where he attempts to find inspiration from how humans do it and build generative model based on common sense reasoning.

SEUNGJUN LEE received the B.S. degree in industrial management engineering from the Hankuk University of Foreign Studies, Yongin, South Korea, in 2021. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea, and the integrated master's and Ph.D. degrees. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include natural language understanding and neural machine translation.

HYEONSEOK MOON received the B.S. degree from the Department of English Language and Literature, Korea University, Seoul, South Korea, in 2021, where he is currently pursuing the Ph.D. degree in computer science and engineering and the integrated master's and Ph.D. degrees. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory Team. His research interests include natural language understanding and neural machine translation.

JUNGSEOB LEE received the B.S. degree from the Department of Information Communication Engineering, Dongguk University, Seoul, South Korea, in 2021. He is currently pursuing the integrated master's and Ph.D. degrees. He is a part of the Natural Language Processing and Artificial Intelligence Laboratory, Korea University. His research interests include machine translation and language modeling.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor at the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.