Towards Standardizing Korean Grammatical Error Correction: Datasets and Annotation

Soyoung Yoon[1], Sungjoon Park[2], Gyuwan Kim[3], Junhee Cho[4], Kihyo Park[5], Gyutae Kim[2], Minjoon Seo[1], Alice Oh[6]

KAIST AI[1], Softly AI[2], University of California, Santa Barbara[3], Cornell University[5], Google[4], KAIST[6]

61 ACL 2023

Try our demo model!

Paper

Introduction

Motivation: Research on Korean Grammatical Error Correction(GEC) is limited. We attribute this problem to the lack of carefully designed evaluation benchmark and resource for Korean GEC.

Contribution:

1. Three large-scale organized **dataset** of Korean GEC
2. KAGAS, an **automatic annotation** & evaluation **system** for Korean GEC
3. Open-sourced baseline error correction **model** based on our data

Dataset

	Kor-Learner	Kor-Native	Kor-Lang8
# Sentence pairs	28,426	17,559	109,559
Avg. token length	14.86	15.22	13.07
# Edits	59,419	29,975	262,833
# Edits / sentence	2.09	1.71	2.40
Avg. tokens per edit	0.97	1.40	0.92
Prop. tokens changed	28.01%	29.37%	39.42%

Kor-Native - data collection by transcription

(Source: Native Speaker)

Corrupted (source)	Corrected (target)
한국에서 뿐만 아니라 세계적으로도 유명해어.	한국에서뿐만 아니라 세계적으로도 유명해요.
이 도시는 공기도 나뿌다.	이 도시는 공기도 나쁘다.
수지씨가 예쁩니까?	수지씨가 예쁩니까?
바다에서 예쁜 조개를 주어가지고 목걸이를 만들어어.	바다에서 예쁜 조개를 주워가지고 목걸이를 만들어요.

Q. Reasonable representative?
A. Followed the distribution of Shin et al., in that the errors are dominated by **word spacing** errors!

Kor-Lang8 - collected by social platform

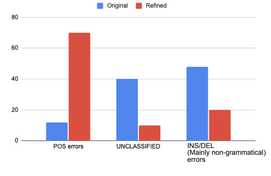
(Source: Second language learner)

Refine & clean up original lang8 dataset by predefined rules

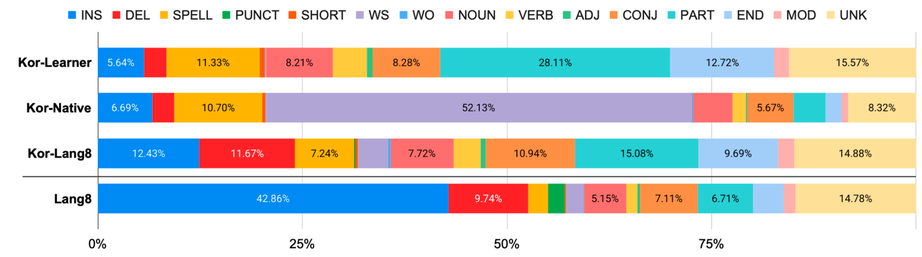
- Remove non-Korean words
- Filter edits by Levenshtein distance
- Discard brackets
- …

	Valid loss	Self-GLEU	GLEU on KoBART	Dataset size
Lang8 (Bef.)	1.53	15.01	19.69	204,130
Kor-Lang8(Aft.)	**0.83**	**19.38**	**28.57**	109,559

* Discard almost half of dataset, with better quality
* Lower loss (stable training), higher GLEU scores



Corrupted (source)	Corrected (target)
지난날 인터넷으로 찾아냈다.	지난달 인터넷 검색으로 찾아냈다.
제가 정식으로 한국말 배워 안했어요.	제가 정식으로 한국어를 배우지 못했어요.
등산하고 바비큐도 하다.	등산도 하고 바비큐도 한다.
고맙다랑 고마워요랑 감사합니다!	고맙다! 고마워요! 감사합니다!



Kor-Learner - essays & annotations by Korean learners and tutors

(Source: Second language learner)

Merge **morpheme**-level edits into word-level edits to make parallel corpora

Apply Korean orthography guidelines 한국어 어문 규범

Kor-Learner contains high proportion of particle errors

Corrupted (source)	Corrected (target)
우리 모두 꿈을 잘 이루어졌으면 좋겠습니다.	우리 모두의 꿈이 잘 이루어졌으면 좋겠습니다.
그 제 꿈이 교수기 도는 것입니다.	제 꿈은 교수가 되는 것입니다.
밥을 먹으로 식당에 가요.	밥을 먹으러 식당에 가요.
그래서 집 관리비가 절약할 수 있어요.	그래서 집 관리비를 절약할 수 있어요.

* original XML file:
과도하ㄴ -> 과도한

```
<LearnerErrorAnnotations>
 <word>
  <p>과도하ㄴ</p>
  <morph from="178" subsequence="1" to="182" wordStart="Start">
   <Proofread pos="NNG">과도</Proofread>
   <ErrorArea type="CNG" />
   <ErrorPattern type="REP" />
   <ErrorLevel type="MDV;POS" />
  </morph>
  <morph from="178" subsequence="2" to="182" wordStart="None">
   <Proofread pos="XSV">하</Proofread>
  </morph>
  <morph from="178" subsequence="4" to="182" wordStart="None">
   <Preserved>ㄴ</Preserved>
  </morph>
 </word>
</LearnerErrorAnnotations>
```

KAGAS

We build an **automatic annotation toolkit** that annotates parallel Korean sentences with error type information given **original & corrected pairs** considering Korean linguistics

	Human annotations	KAGAS
Schema	Different by datasets. Cannot compare	Provides a unified schema
Decisions	Random Differs by annotators	Deterministic More trustworthy
Time/price cost	Experts needed. Expensive	No cost, can instantly get output

Motivation: Annotation by humans have disadvantages

What is different for Korean?

- Classifying error types by ~~word~~ morpheme-level

- Morpheme-level **del**etion
  - 소풍(NNG) + 을(JKO) =>소풍(NNG)
  - 을(JKO) is deleted, thus labeled as PART(JKO is grouped to PART)
- Morpheme-level **in**sertion
  - 유학(NNG) => 유학(NNG) + 러(NNP)
  - 러(NNP) inserted, thus labeled as Noun(NNP -> PART)

```
Orig: 그는 방을 저장한 다음에 갔다 학교에 .
Corr: 그는 방을 정리한 다음에 학교에 갔다 .
KAGAS output:
S 그는 방을 저장한 다음에 갔다 학교에 .
A 2 3|||VERB|||정리한|||REQUIRED|||-NONE-|||0
A 4 6|||WO|||학교에 갔다|||REQUIRED|||-NONE-|||0
```

```
Orig: He to school went after saving his room .
Corr: He went to school after cleaning his room .
KAGAS output:
S He to school went after saving his room .
A 4 5|||VERB||cleaning|||REQUIRED|||-NONE-|||0
A 1 3|||WO|||went to school|||REQUIRED|||-NONE-|||0
```

- No PREP, but PART for Korean

School->학교, To->에, 학교에->School-To(not To-School)

"To" is not **prep**osition, but rather a postpositional **part**icle

- About INS/DEL edits.

Korean is a discourse-oriented language. One can omit the subject or object in a sentence depending on the previous context, which is grammatically correct in most cases.

Correlation (kappa) scores		Value	Reliability
Fleiss' Krippendorff'		**0.4386** **0.4392**	moderate moderate
Cohen's (pairwise)	Ann.1&2	0.5976	moderate
	Ann.1&3	0.4426	moderate
	Ann.2&3	0.3566	fair
	Average	**0.4656**	moderate

Correlation scores of 3 GEC experts

Acceptance rate of KAGAS by 3 GEC experts

Dataset	Coverage	Overall acceptance rate			
		Total	Evaluator 1	Evaluator 2	Evaluator 3
Kor-Learner	81.56%	87.34% ± 5.49%P	84.32% ± 10.70%P	87.81% ± 8.26%P	89.88% ± 7.44%P
Kor-Native	90.92%	93.93% ± 2.18%P	92.92% ± 4.14%P	93.99% ± 2.82%P	94.87% ± 3.60%P
Kor-Lang8	82.52%	87.06% ± 4.67%P	84.72% ± 8.88%P	86.98% ± 7.02%P	89.48% ± 6.97%P

Error Code	Description & Acceptance Rate (%)	Example
INS	A word is inserted. 100.00% ± 0.00%P	Original: 고등학교 때 어긴 경험 Corrected: 고등학교 때 **규칙을** 어긴 경험 Translation: Experience to break a **rule** in high school
DEL	A word is deleted. 100.00% ± 0.00%P	Original: 전쟁 **끝** 직후 장군들은 사형을 선고 받았다. Corrected: 전쟁 직후 장군들은 사형을 선고 받았다. Translation: After the war, the generals are sentenced to death.
WS	Spacing between words is changed. 100.00% ± 0.00%P	Original: **이옷은** 더러워요. Corrected: **이 옷은** 더러워요. Translation: This **cloth** is dirty.
WO	The order of words is changed. 97.44% ± 3.51%P	Original: 저는 **더 한국어를** 배우고 싶어요. Corrected: 저는 **한국어를 더** 배우고 싶어요. Translation: I want to learn **Korean further**.
SPELL	Spelling error 97.44% ± 3.51%P	Original: 파티에서 우리는 춤을 **쳐요.** Corrected: 파티에서 우리는 춤을 **춰요.** Translation: We **dance** at the party.
PUNCT	Punctuation error 98.72% ± 2.50%P	Original: 1993년 **의** 겨울의 일이였다. Corrected: 1993년 , 겨울의 일이였다. Translation: It was 1993, a happening in winter.
SHORT	An edit that does not change the structure of morphemes. 73.08% ± 9.84%P	Original: 한국어는 저한테 너무 어려운 **언어이었어요.** Corrected: 한국어는 저한테 너무 어려운 **언어였어요.** Translation: Korean **Language** was too difficult to me.
VERB	An error on verb 79.49% ± 8.96%P	Original: 어제 친구에게 편지를 **짰어요.** Corrected: 어제 친구에게 편지를 **썼어요.** Translation: I **wrote** a letter to my friend yesterday.
ADJ	An error on adjective 73.08% ± 9.84%P	Original: **진한** 친구 Corrected: **친한** 친구 Translation: A **close** friend.
NOUN	An error on noun 75.64% ± 9.53%P	Original: 나중에 기회가 있을 때 한국에 **유학러** 가고 싶습니다. Corrected: 나중에 기회가 있을 때 한국에 **유학** 가고 싶습니다. Translation: I want to study **abroad** in Korea in the future.
PART	An error on particle 97.44% ± 3.51%P	Original: **하와이에서** 사는 우리 사촌 Corrected: **하와이에** 사는 우리 사촌 Translation: My cousin living **in Hawaii**
END	An error on ending 87.18% ± 7.42%P	Original: 오래 **기다려요.** Corrected: 오래 **기다렸어요.** Translation: I **waited** for a long time.
MOD	An error on modifier 89.74% ± 6.73%P	Original: 점심이 **나무** 작은 나머지 배고팠어요. Corrected: 점심이 **너무** 작은 나머지 배고팠어요. Translation: I was hungry because I had **such** a small launch.
CONJ	An error on conjugation 43.59% ± 11.00%P	Original: 오늘은 머리를 **잘라에** 갔다. Corrected: 오늘은 머리를 **자르러** 갔다. Translation: I went to a barber to get my hair **cut** today.

Full list of KAGAS Error Types with examples

Baseline Models

Trained our data on top of SKT-AI/ KoBART and compared with Hanspell, a widely used GEC system based on statistical modeling.

Evaluation dataset ->

Neural Models > Hanspell

Model variants ↓	Kor-Learner				Kor-Native				Kor-Lang8				Kor-Union				Gen. time
	GLEU	M^2			GLEU	M^2			GLEU	M^2			GLEU	M^2			
		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$		Pre.	Rec.	$F_{0.5}$	
Self-Scores	25.54	1	0	0	25.71	1	0	0	20.01	1	0	0	21.66	1	0	0	
Hanspell	30.36	29.45	5.33	15.46	57.08	81.93	47.36	71.50	22.94	29.18	8.74	19.88	28.82	37.34	11.58	25.85	189.69
KoBART	**45.06**	43.35	**24.54**	37.58	**67.24**	75.34	**55.95**	70.45	28.48	**37.56**	11.62	25.93	**33.70**	**44.75**	**14.64**	**31.70**	38.25
KoBART + Kor-Union	42.66	**53.51**	21.18	**41.00**	59.71	**85.47**	47.38	**73.63**	**28.65**	37.46	**12.00**	**26.78**					37.51

Scores of KoBART are evenly distributed for all error types than Hanspell

		FULL	STD	INS	DEL	WS	WO	SPELL	PUN.	SHO.	VERB	ADJ	NOUN	PART	END	MOD	CONJ	UNK
GLEU	Hanspell	28.82	9.19	**21.89**	13.57	**46.91**	7.51	31.73	16.81	31.06	19.94	18.31	19.67	17.14	19.17	17.41	19.79	**21.15**
> 20	KoBART	33.70	6.02	**23.67**	18.22	**34.83**	13.76	**36.10**	**20.29**	**33.42**	**27.96**	**29.27**	**25.09**	**27.15**	**27.72**	**23.68**	**25.95**	**26.44**
Prec.	Hanspell	37.34	12.22	37.49	33.51	**71.98**	25.00	**48.21**	30.61	**46.70**	30.32	28.42	29.94	24.37	27.12	31.41	32.10	34.07
> 40	KoBART	44.75	6.78	**40.58**	**45.64**	**63.26**	37.89	**52.16**	**42.35**	**58.52**	**46.45**	**49.16**	**43.76**	**49.35**	**47.51**	**46.67**	**41.30**	**42.69**
Recall	Hanspell	11.58	9.00	**10.10**	7.40	**39.30**	5.26	**21.25**	6.14	**15.48**	6.62	6.12	8.14	5.01	5.64	7.08	8.60	9.69
> 10	KoBART	14.64	4.09	9.47	**12.05**	**21.48**	9.04	**22.54**	**10.06**	**19.61**	**13.29**	**13.92**	**12.46**	**14.30**	**13.23**	**12.70**	**12.34**	**12.68**
$F_{0.5}$	Hanspell	25.85	12.58	24.31	19.64	**61.72**	14.29	**38.46**	17.03	**33.28**	17.67	16.43	19.50	13.75	15.40	18.61	20.76	22.66
> 30	KoBART	31.70	6.46	24.47	29.28	**45.48**	23.11	**41.30**	25.78	**41.84**	**30.98**	**32.61**	29.11	**33.10**	**31.28**	**30.38**	28.09	28.97