Rethinking Annotation: Can Language Learners Contribute?

Haneul Yoo[1], **Rifki Afina Putri**[1], **Changyoon Lee**[1], **Youngin Lee**[1],
So-Yeon Ahn[1], **Dongyeop Kang**[2], **Alice Oh**[1]
[1]KAIST, South Korea, [2]University of Minnesota, USA
{haneul.yoo, rifkiaputri, cyoon47, conviette}@kaist.ac.kr,
ahnsoyeon@kaist.ac.kr, dongyeop@umn.edu, alice.oh@kaist.edu

Abstract

Researchers have traditionally recruited native speakers to provide annotations for widely used benchmark datasets. However, there are languages for which recruiting native speakers can be difficult, and it would help to find learners of those languages to annotate the data. In this paper, we investigate whether language learners can contribute annotations to benchmark datasets. In a carefully controlled annotation experiment, we recruit 36 language learners, provide two types of additional resources (dictionaries and machine-translated sentences), and perform mini-tests to measure their language proficiency. We target three languages, English, Korean, and Indonesian, and the four NLP tasks of sentiment analysis, natural language inference, named entity recognition, and machine reading comprehension. We find that language learners, especially those with intermediate or advanced levels of language proficiency, are able to provide fairly accurate labels with the help of additional resources. Moreover, we show that data annotation improves learners' language proficiency in terms of vocabulary and grammar. One implication of our findings is that broadening the annotation task to include language learners can open up the opportunity to build benchmark datasets for languages for which it is difficult to recruit native speakers.

1 Introduction

Data annotation is important, and in NLP, it has been customary to recruit native speakers of the target languages, even though it is difficult to recruit native speakers for many languages. Meanwhile, there are many people learning another language, for instance, Duolingo claims that 1.8 billion people are learning a foreign language using their app. [1]

In this paper, we examine whether language learners can annotate data as well as native speakers and whether their annotations can be used to train language models. We explore this question with five control variables that may affect the quality of language learner annotations. These are the language, task, learners' language proficiency, difficulty of the annotation questions, and additional resources that learners can consult. We recruited learners at various levels of proficiency in English (high-resource), Korean (mid-resource), and Indonesian (low-resource). They annotated data on four tasks, sentiment analysis (SA), natural language inference (NLI), named entity recognition (NER), and machine reading comprehension (MRC). We ask three levels of learners to complete multiple sessions of data annotation given with the help of a dictionary or machine-translated texts.

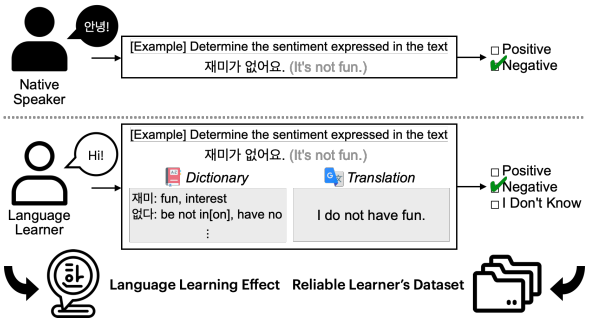


Figure 1: Recruiting language learners in NLP data annotation. They can be assisted by dictionaries or MT systems.

Our major findings, both in terms of the quality and learning effect of learners' annotations, are summarized as follows: We measure the degree of inter-annotator agreement between learners and ground truth labels, and show that *language learners can annotate data at a fairly accurate level*, especially for the simpler tasks of SA and NER, and for easy- to medium-level questions. Language learners consulting dictionaries generate more accurate labels than learners consulting machine-translated sentences. Language

[1] https://www.duolingo.com/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14714–14733

models trained on data generated from the distribution of the learners' annotations achieved performance comparable to those of models trained on ground truth labels, demonstrating the efficacy of learner-annotated data.

We also observe that *learners' language proficiency in vocabulary and grammar tends to improve* as they carry out the annotation tasks. We measure their proficiency by conducting pre- and post-tests before and after the annotation. Learners perceive that their language proficiency improved during data annotation, and most were willing to re-participate in the process.

We hope this paper allows researchers to question the necessity of recruiting native speakers for data annotation and call on other NLP researchers carefully to consider the criteria by which to recruit crowdworkers for data annotation carefully.

2 Related Work

We can group annotators of NLP datasets into language learners, non-speakers, and non-experts. Language learners are people who are learning the target language, while non-speakers are those who have never learned the target language. Non-experts are people who have no expertise in NLP tasks or data annotations. We look at previous work with these three annotator groups.

Language Learner Annotation. There are several tools for both language learning and crowdsourcing that create linguistic resources. The early motivation of Duolingo was to translate the web with language learners (von Ahn, 2013). Hladká et al. (2014) introduced a pilot experiment on Czech, the aim of which was both data annotation and the teaching of grammar. Sangati et al. (2015) proposed a web-based platform similar to that of Duolingo that undertakes POS tagging with grammar exercises through interactions between a teacher's validation and students' annotations. Nicolas et al. (2020) employed language learners to extend existing language resources (ConceptNet (Liu and Singh, 2004)), showing that this method also has educational values. However, they did not explicitly mention the details of their experimental settings, including the number of participants, and there was no study that recruited and employed language learners in NLP tasks with a comprehensive empirical analysis of diverse factors.

Non-speaker Annotation. A recent study employed non-speakers on specific NLP tasks and provided tools for non-speaker annotators, but that study mainly focused on easy tasks such as NER and binary classification tasks. Tsygankova et al. (2021) employed non-speakers as annotators to build a NER dataset and model for Indonesian, Russian, and Hindi and compared their performances with those of fluent speakers'. The non-speakers produced meaningful results for NER in Indonesian on a combination of an easy task and an easy language written in the Latin alphabet with simple grammar. Mayhew et al. (2020); Kreutzer et al. (2022) also employed non-speakers for some easy tasks such as NER along with native or fluent speakers. Despite these efforts, it remains unclear as to whether non-speakers can undertake annotation on more complex tasks such as MRC with a paragraph to read, and NLI, requiring a comprehensive understanding of the premise and hypothesis sentences to infer the connection between the sentences correctly.

Hermjakob et al. (2018); Mayhew and Roth (2018); Lin et al. (2018); Costello et al. (2020) devised assisting tools for non-speaker annotation, providing English translation, romanization, dictionary matching, and grammar-related descriptions. We expect that English translation and dictionary matching may also be helpful to language learners and adopt the same setup. However, neither romanization nor grammar-related descriptions may help the learners because they already have some background knowledge of the target language, unlike the non-speakers.

Non-expert Annotation. Snow et al. (2008) suggested using a collection of non-expert annotations rather than expensive expert annotations. They analyzed and compared those two types of annotations on several NLP tasks. Only relatively few non-expert annotations are necessary to equal the performance of an expert annotator for certain simple tasks. Madge et al. (2019) suggest the training of non-expert annotators via progression in a language annotation game considering the linguistic ability of crowdworkers and the readability level of documents.

3 Study Design

This section describes how we carefully design our controlled experiments with diverse factors that may affect the quality of learners' annotations and

the learning effect.

3.1 Control Variables

Table 2 shows a summary of the different control variables considered in our experiments with the corresponding values. We should take these control variables into account when simulating learners' annotations in real-world scenarios and use diverse combinations of them. We set the major control variables based on previous work on NLP data annotation (Joshi et al., 2020; Wang et al., 2018; Lin et al., 2018) and language learning (Lee and Muncie, 2006; Crossley et al., 2008; Shieh and Freiermuth, 2010).

Language Selection. We choose three target languages, English (EN), Korean (KO), and Indonesian (ID), based on the availability of gold-label data, the availability of native speakers to evaluate, and the difficulty of the language. English is the highest-resource language, while Korean and Indonesian are mid- to low-resource languages, respectively (Joshi et al., 2020). Korean uses its own alphabet, while Indonesian adopts the Latin alphabet. The Foreign Service Institute (FSI) [2] categorizes languages into five categories based on the amount of time it takes to learn them considering several variables, including grammar, vocabulary, pronunciation, writing system, idiomatic expressions, distance from English, dialects, and learning resources. According to the FSI ranking, Indonesian is in category 2, requiring around 36 weeks or 900 class hours, Korean is in category 4, requiring 88 weeks or 2200 class hours to reach B2/C1 level in CEFR, and English is in category 0.

Task and Data. We choose four tasks from each common task type in the GLUE benchmark (Wang et al., 2018): sentiment analysis (SA) for single sentence classification, natural language inference (NLI) for sentence pair classification, named entity recognition (NER) for sequence tagging, and machine reading comprehension (MRC) for span prediction. Table 1 presents a list of the datasets used in our study. SA has two options (`positive` and `negative`), and NLI has three options (`entailment`, `neutral`, and `contradict`) for all languages. The NER datasets have different categories of named entities among the languages, while all languages have `person` and `location` entities.

Participant Selection. We adopt and revise the CEFR [3] criteria to categorize learners into three levels: basic (A1-A2), intermediate (B1-B2), and advanced (C1-C2). Table 3 shows our recruiting criteria with respect to language fluency. We do not request official test scores for basic-level learners, as they may not have taken official language proficiency tests. We assign the learners at each level to annotate questions to facilitate majority voting among three responses from different levels of participants. All annotators in our experiments are non-experts in NLP data annotations, and three annotators are allocated to each task and each additional resource. Participants are asked to do two tasks: SA and MRC, or NER and NLI. The study involved participants with ages ranging from 19 to 44 (average 31.5, median 24) at the time of the experiment. They are primarily undergraduate or graduate students, with some office workers and unemployed individuals.

Additional Resources. Lin et al. (2018) observed that additional resources such as dictionary matching or English translation may assist non-speakers with annotation tasks. We divide the participants into two groups with the additional resources at their disposal, in this case a dictionary and translations provided by a commercial MT system. We only provide texts in the target language and ask participants to consult online or offline dictionaries if they need any help in the dictionary setting. Otherwise, we provide both the texts in the target language and corresponding translations created by the Google Translate API on our website and ask the participants not to use any other external resources.

Annotation Sample Selection. We randomly sample 120 annotation samples for each task from the source datasets and categorize them into five groups based on their difficulty level. The sentence-level difficulty score is calculated using a macro average of several linguistic features from Coh-Metrix (Graesser et al., 2004), a metric for calculating the coherence and cohesion of texts. The linguistic features that we use in our experiment are the *lexical diversity*, *syntactic complexity*, and *descriptive measure*. Lexical diversity is computed by the type-token ratio, syntactic complexity is computed according to the number of conjunction

[2] https://www.state.gov/foreign-language-training/

[3] Common European Framework of Reference for Languages (https://www.coe.int/en/web/common-european-framework-reference-languages)

	SA	NLI	NER	MRC
EN	SST2 (Socher et al., 2013)	SNLI (Young et al., 2014)	CoNLL++ (Tjong Kim Sang and De Meulder, 2003)	TyDiQA (Clark et al., 2020)
KO	NSMC (Park, 2016)	KLUE (Park et al., 2021)	KLUE (Park et al., 2021)	TyDiQA (Clark et al., 2020)
ID	IndoLEM (Koto et al., 2020)	IndoNLI (Mahendra et al., 2021)	NERP (Wilie et al., 2020)	TyDiQA (Clark et al., 2020)

Table 1: Source dataset for each language and task.

Control Variables	Values
Language	EN, KO, ID
Task	SA, NLI, NER, MRC
Learner Fluency	Basic, Intermediate, Advanced
Question Difficulty	Very easy, ⋯, Very hard
Additional Resources	Dictionary, Translation

Table 2: Control variables in our experiments.

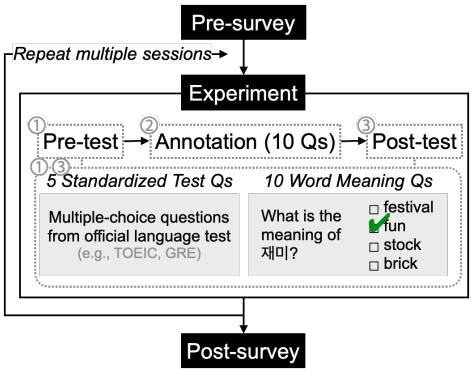


Figure 2: High-level flowchart of our experiments.

words, and descriptive measure is computed by the sentence character length, the number of words, and the mean of the number of word syllables. We add additional metrics for MRC tasks that contain a paragraph, in this case the number of sentences in the paragraph, the character length of the answer span, and the number of unique answers. The paragraph-level difficulty score is calculated by taking the average of the sentence-level scores in the paragraph.

Test Question Selection. Pre- and post-tests are used, consisting of five questions from official language proficiency tests and ten questions asking about the meanings of words appearing in annotation samples that they will solve in the same session. Standardized test questions explore whether participating in the annotation improves the learners' overall language proficiency over several days, while word meaning questions aim to inspect whether participating in the annotation helps them learn some vocabulary.

We use TOPIK[8] for Korean, UKBI[9] and BIPA[10] for Indonesian, and TOEIC[11] and GRE[12] for English. We chose nouns and verbs from annotation questions and created multiple-choice questions whose answers are the nouns or the verbs in the annotation questions.

3.2 Workflow

Step 1: Pre-survey As shown in Figure 2, we use a survey to ask participants about their self-rated language fluency, language background, and learning experience before the main experiments. We describe the CEFR criteria and ask participants to self-evaluate their language proficiency in general, colloquial, and formal texts and choose which of the colloquial and formal texts they are more familiar with.

Step 2: Experiment Our experiments consist of a series of multiple sessions over six days. Each session consists of three steps, and we ask participants to do two sessions per task per day and

[4]Test Of English as a Foreign Language (https://www.ets.org/toefl)

[5]Oral Proficiency Interview (https://www.actfl.org/assessment-research-and-development/actfl-assessments/actfl-postsecondary-assessments/oral-proficiency-interview-opi)

[6]Foreign Language EXamination (https://www.kotga.or.kr/sub/sub03_01.php)

[7]Tes Bahasa Indonesia sebagai Bahasa Asing (https://lbifib.ui.ac.id/archives/105)

[8]Test Of Proficiency In Korean (https://www.topik.go.kr/)

[9]Uji Kemahiran Berbahasa Indonesia (https://ukbi.kemdikbud.go.id/)

[10]Bahasa Indonesia untuk Penutur Asing (https://bipa.ut.ac.id/)

[11]Test Of English for International Communication (https://www.ets.org/toeic)

[12]Graduate Record Examination (https://www.ets.org/gre)

	Basic	Intermediate	Advanced
EN	Self report A	TOEFL [4] 57-109	TOEFL $\geq$ 110
KO	Learning experience < 1 yr & Self report A	TOPIK Level 2-4	TOPIK $\geq$ Level 5
ID	Learning experience < 1 yr & Self report A	OPI [5] $\leq$ IH $\|\|$ FLEX [6] $\approx$ 600	OPI $\geq$ AL $\|\|$ TIBA [7] $\geq$ 4

Table 3: Learner level criteria

		Accuracy	Inter-Annotator Agreement	Time (min)
Native Speakers	-	$8.53_{\pm 0.09}$	$0.77_{\pm 0.02}$	$4.07_{\pm 0.78}$
Language Learners	Dictionary	$7.72_{\pm 0.09}$	$0.70_{\pm 0.01}$	$6.92_{\pm 0.70}$
	Translation	$7.31_{\pm 0.09}$	$0.67_{\pm 0.01}$	$6.49_{\pm 0.36}$

Table 4: Annotation comparison between native speakers and learners (with dictionary and translation settings). Accuracy means the number of correct questions compared to the ground truth labels out of 10. Inter-Annotator agreement means pairwise F1-score. Time means how long annotating 10 samples takes in minutes.

repeat this for six consecutive days. Before starting the main experiment, we provide a pilot session to check whether the participants fully understand our instructions. All of the experimental processes are done on our research website, and we measure the time spent by the participants on each step.

Step 2.1: Pre-test Participants solve 15 test questions to check their language proficiency level. All test questions are multiple-choice types and include the "*I don't know*" option.

Step 2.2: Annotation Participants annotate ten questions with the help of the additional resources assigned.

Step 2.3: Post-test After completing the annotation, participants solve the same 15 test questions they solved in the pre-test. This step investigates whether data annotation has any learning effect.

Step 3: Post-survey After the experiments, participants complete a post-survey about their thoughts on annotation and self-rated language proficiency. They answer the questions below for each task on a five-point Likert scale from "*strongly disagree*" to "*strongly agree*".

4 Experimental Results

We discuss the results of our experiments with respect to two research questions:

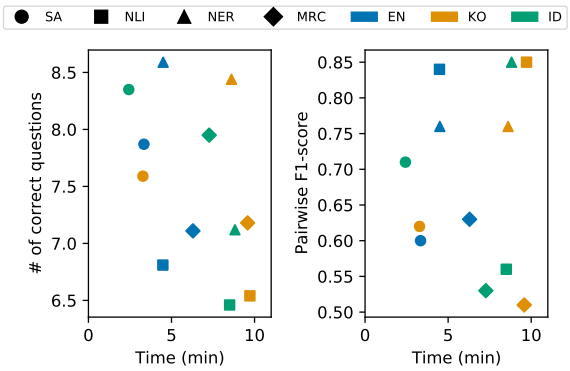
1. Can we obtain a reliable dataset from learners' annotations? Which design setting would be most helpful? We answer this question via quality assessment (§4.1), training simulation (§4.2), and error analysis (§5.1).
2. Do learners improve their language proficiency while annotating the NLP tasks (§5.2)?

All findings we discuss in this section were shown to be statistically significant at p level of < 0.05 using ANOVA. Specifically, comparisons for annotation accuracy, annotation time, and survey responses were analyzed with four-way ANOVA over the four between-subject factors of task, language, additional resources, and learner level. Comparisons between pre-test and post-test results were done with a mixed two-way ANOVA with learner level and additional resources as between-subject factors. Pairwise t-tests were conducted for all factors with Bonferroni corrections.

4.1 Annotation Quality

Accuracy and Agreement. Table 4 shows the results of annotations generated by language learners compared to native speakers. Language learners made correct annotations to 7.48 questions among 10 questions on average [13], taking 6.68 minutes. They generated 1.05 less accurate labels and took 2.6 minutes longer time than the native speakers. Learners assisted by dictionaries can produce more reliable labels than learners using MT system. Meanwhile, majority voting among native speakers generated 19 incorrect labels out of 120 questions, compared to learners' 21.5 incorrect labels (Table 11 in Appendix). This shows that language learners' annotations can be aggregated by majority voting to be nearly as accurate as those of native speakers.

[13] Annotation accuracy was computed by a weighted averaged F1 score compared to the ground truth label on NER and MRC. The average of the weighted-averaged F1 score was used for some samples in MRC with multi-choice answers.



(a) Annotation accuracy (b) Inter-annotator agreement

Figure 3: Task difficulty according to language and task.

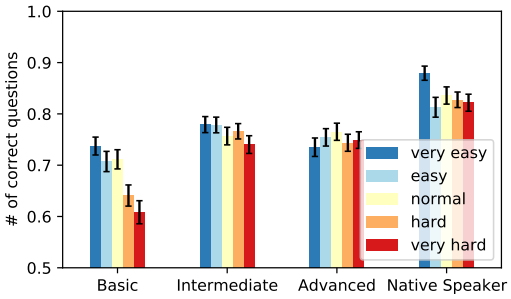


Figure 4: Annotation accuracy on question difficulty and language fluency.

Languages and Tasks. Figure 3 (a) and (b) show the task difficulty with respect to time versus annotation accuracy and inter-annotator agreement, respectively. SA and NER are easier for language learners than NLI and MRC, considering both accuracy and time. MRC, which requires paragraph comprehension, unlike sentence-level tasks, may be difficult for learners. Nonetheless, they achieved high accuracy, and most of their answer spans overlapped with the ground truth answers. Detailed results and further analysis of the outcomes in Figure 3 can be found in Appendix B.

We measure inter-annotator agreement using the pairwise F1 scores. Table 10 (b) shows the level of agreement and the standard error for each language and task. Both NLI and NER show high agreement, while the token-based task MRC shows relatively low agreement compared to the other tasks.

Korean SA shows low agreement, most likely due to some noisy samples in the NSMC dataset. The NSMC dataset is a movie review dataset whose negative labels come from the reviews with ratings of 1-4, and where the positive labels come from those with ratings of 9-10, respectively. This dataset contains noisy samples whose gold labels are unreliable or whose labels cannot be determined only with the text, requiring some metadata.

MRC in Korean shows low agreement, and we assume this stems from the fact that Korean is a morpheme-based language while the others use word-based tokenization. The F1 score was computed based on the corresponding word overlaps in both English and Indonesian. Korean uses character-based overlap, which is stricter. It may be more complicated for annotators to clearly distinguish the answer span at the character level rather than at the word level.

Language Proficiency and Question Difficulty. Figure 4 shows the percentage and the standard error of obtaining a correct answer for each question difficulty and learner fluency. Both intermediate and advanced learners show similar levels of accuracy regardless of question difficulty level, while basic-level learners tend to fail on complex questions. The mean number of correct questions out of 10 increases to 7.66 without basic-level learners. This implies that the intermediate level is sufficient to understand the sentences in general NLP datasets and suggests the feasibility of recruiting learners as annotators in place of native speakers, especially on easy-to-medium tasks and questions.

4.2 Training Simulation with Learners' Annotations

In order to show the reliability of learners' annotations used as training labels for language models, we compare the performance of models trained on learners' annotations across SA and NLI to the models trained on native speakers' annotations. Because we only have a small number of learners' annotations, we generate synthetic data following the distribution of learners' annotations. We randomly select 10K samples from the training data of the original datasets and change the labels into the generated synthetic labels. We aggregate learners' annotations using a majority vote. We ran the Shapiro-Wilk test and found that the distribution of labels is Gaussian (p-value < 0.05). We then fit the probability distributions of labels for each class and generate synthetic labels for existing NLP datasets based on those distributions. The same process is used to build synthetic data representing native speakers' annotations. We set two baselines as the upper and lower bounds of LMs: models trained on the original ground truth labels (Ground Truth) and models trained on machine-translated texts of

		SA			NLI		
		EN	KO	ID	EN	KO	ID
Ground Truth	-	$89.56_{\pm 1.11}$	$85.29_{\pm 0.79}$	$97.20_{\pm 0.86}$	$79.05_{\pm 1.44}$	$79.00_{\pm 2.48}$	$68.20_{\pm 1.32}$
MT Dataset	-	$79.25_{\pm 1.25}$	$75.27_{\pm 1.33}$	$87.19_{\pm 1.39}$	$56.78_{\pm 2.26}$	$47.06_{\pm 1.26}$	$52.35_{\pm 1.35}$
Native Speakers	-	$87.59_{\pm 1.30}$	$89.18_{\pm 1.62}$	$94.18_{\pm 0.31}$	$71.86_{\pm 1.43}$	$74.09_{\pm 1.41}$	$67.21_{\pm 1.23}$
Language Learners	All	$89.09_{\pm 1.87}$	$89.26_{\pm 1.41}$	$94.26_{\pm 0.73}$	$72.16_{\pm 1.91}$	$71.82_{\pm 1.27}$	$70.39_{\pm 2.17}$
	Dictionary	$86.64_{\pm 0.40}$	$87.61_{\pm 1.06}$	$92.61_{\pm 1.44}$	$70.40_{\pm 1.09}$	$74.22_{\pm 1.98}$	$66.70_{\pm 1.48}$
	Translation	$85.39_{\pm 1.14}$	$87.47_{\pm 1.65}$	$92.47_{\pm 1.46}$	$74.69_{\pm 1.63}$	$73.03_{\pm 1.02}$	$69.84_{\pm 2.49}$

Table 5: Training simulation of annotations by native speakers and learners. BERT-based models are trained on labels generated or synthesized by each group. We provide the upper and lower bounds on the performances based on ground-truth labels and translations, respectively.

Task	Top-3 Failure Reasons
SA	• Unreliable gold label • Lack of background information • Ungrammatical sentence [14]
NLI	• Task ambiguity • Unreliable gold label • Domain-specific genre and expression
MRC	• Culturally-nuanced expression • Ambiguous questions with multiple answers • Low overlaps in answer span

Table 6: Main failure reasons on each task

other languages (MT Dataset).

We fine-tuned $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019), $\text{KLUE-BERT}_{\text{BASE}}$ (Park et al., 2021), and $\text{IndoBERT}_{\text{BASE}}$ (Wilie et al., 2020) for English, Korean, and Indonesian, respectively. Table 5 shows the experimental results of the LMs trained on different synthetic labels, averaged for each language. Ground Truth indicates LMs trained on the original label, which was annotated by native speakers and merged into one by majority vote. Models trained on synthetic labels representing learners' annotations significantly outperformed the MT Dataset. This implies that building datasets with learners' annotation can produce more reliable labels than the baseline method of using machine-translated high-resource language datasets.

5 Discussion

5.1 Qualitative Analysis on Learners' Annotation

We analyze the annotation result of each sample, especially the samples on which learners or native speakers failed, i.e., those that were incorrectly labeled or for which "*I don't know*" was selected as the answer. Table 6 shows the main failure reasons why learners failed to make correct annotations on each task for the samples that at most one learner correctly labeled. The number of samples for which all learners failed ranges from zero to three for all tasks, except for NER, where no sample was incorrectly predicted by all learners; i.e., there was at least one learner who answered correctly for each question for all 120 samples.

We found that the incorrectly labeled samples in SA mostly occurred due to the unreliable gold label in the dataset. With regard to NLI, all incorrect samples resulted from ambiguities in the task itself. Some NLI and MRC samples are tricky for learners in that they can create correct labels only when they fully understand both the hypothesis and the premise or both the context and the question. Fluency in English may affect failures by Indonesian learners in the translation setting, considering that the provided translations were in English. *Very difficult* examples in MRC occasionally include difficult and culturally-nuanced phrases and require background knowledge, which can be difficult for learners.

A detailed explanation of the failure reason analyses results is provided in Table 20 in the Appendix. For instance, a missing period between two short sentences, *스토리가 어려움 (The story is difficult)* and *볼만함 ([but it's] worth watching.)*, in Table 20 (a) leads to misunderstandings among learners. Also, an ambiguity of NLI whether "*people*" and "*some people*" in premise (*People standing at street corner in France.*) and hypothesis (*Some people are taking a tour of the factory.*) are indicating the same leads all learners and native

[14] e.g., missing period, missing spacing and blank, nominalization, and use of slang

	Basic	Intermediate	Advanced
pre-test	$2.72_{\pm 0.09}$	$3.68_{\pm 0.08}$	$3.99_{\pm 0.07}$
post-test	$2.76_{\pm 0.09}$	$3.62_{\pm 0.08}$	$4.01_{\pm 0.06}$

(a) Number of correct standardized test questions out of 5

	Basic	Intermediate	Advanced
pre-test	$7.29_{\pm 0.12}$	$8.93_{\pm 0.08}$	$9.32_{\pm 0.06}$
post-test	$8.41_{\pm 0.11}$	$9.27_{\pm 0.07}$	$9.42_{\pm 0.06}$

(b) Number of correct word meaning questions out of 10

Table 7: Pre-/post-test score in the same session

	Basic	Intermediate	Advanced
1st	$3.23_{\pm 0.02}$	$3.26_{\pm 0.02}$	$3.30_{\pm 0.02}$
last	$3.43_{\pm 0.03}$	$3.46_{\pm 0.03}$	$3.53_{\pm 0.02}$

(a) Number of correct standardized test questions out of 5

	Basic	Intermediate	Advanced
1st	$8.20_{\pm 0.03}$	$8.23_{\pm 0.03}$	$8.30_{\pm 0.03}$
last	$8.91_{\pm 0.03}$	$8.95_{\pm 0.03}$	$9.00_{\pm 0.03}$

(b) Number of correct word meaning questions out of 10

Table 8: Pre-test score of the first and the last session

speakers to get confused between `neutral` and `contradiction`, which is an ambiguity of NLI itself (Table 20 (b)).

5.2 Learning Effect

Standardized Test Questions. We compared pre- and post-test scores for the standardized questions in Table 7 (a). There was no significant difference, implying that annotating several questions had little impact on learning grammar, structure, or general language skills in the short term.

Word Meaning Questions. Table 7 (b) shows the scores of the pre-/post-tests on the word meaning questions out of 10 questions. The learning effect on vocabulary was maximized with beginner-level learners. Both intermediate and advanced learners achieved a mean score of about 9 out of 10 on the pre-test, implying that words used in the data annotation sentences were accessible and understandable enough for them.

Long-term Learning Effect. The pre-test score for the last session is higher than that for the first session by about 4% and 7% each on both standardized test questions and word meaning questions, respectively (Table 8). The increase in the standardized test question scores implies learners' improvement on general language proficiency factors, including structure and grammar. Also, we can surmise that the vocabulary or expressions used in the NLP datasets are primarily redundant and repetitive, considering that only a few sessions can lead to an increase in pre-test scores.

5.3 Concerns about Learners' Annotation in Low-resource Languages

This paper suggests recruiting language learners as crowdworkers in data annotation in low-resourced languages by proving the quality of learners' labels. There are clearly many low-resource languages for which the absolute number of native speakers is exceptionally small compared to learners or for which it is almost impossible to find native speakers in the locations where NLP research is active. For instance, we can think of endangered languages such as Irish, which has no monolingual native speaker and extremely few daily-using L1 speakers (73K) but more than 1M learners. We can also count local languages, such as Sundanese in Indonesia and Jejueo in Korea, that are spoken by the elderly in the community, with the younger speakers who are not fluent but who are much more accessible to the researchers for annotation.

We may use either MT systems such as Google Translate considering that it supports 133 languages including several low-resource languages [15] or dictionaries for extremely low-resource languages such as Ojibwe People's Dictionary [16]. For low-resource languages, it is necessary to scrape together whatever resources are accessible, regardless of whether these are (incomplete) dictionaries, semi-fluent speakers, and/or anyone willing to learn and annotate in that language.

6 Conclusion

This study provides interesting results both for the actual dataset annotation as well as understanding the non-native speakers' annotation capabilities. We show (1) labels provided by language learners are nearly as accurate, especially for easier tasks, (2) with additional experiments of aggregating their labels, learners' are almost on par with native speakers, and (3) language models trained

[15] https://github.com/RichardLitt/low-resource-languages

[16] https://ojibwe.lib.umn.edu/

		Time / Session (min)	Expected Hourly Wage
Native Speakers	-	$8.08_{\pm 0.89}$	KRW 9,282
Language Learners	Dictionary	$14.76_{\pm 1.29}$	KRW 20,325
	Translation	$13.20_{\pm 0.73}$	KRW 22,727

Table 9: Expected hourly wage of each experiment. All wages are over the minimum wage in the Republic of Korea (KRW 9,160).

on learners' less accurate labels achieved 94.44% of ground truth performance.

By showing that NLP annotation does not require finding native speakers, we show the possibility of broadening NLP research for more languages, as it is very challenging to recruit native speakers for many languages. Requiring native speakers for annotation can mean traveling to remote locations and working with an older, less-technology-savvy population. We show that it is possible to work with language learners to hurdle geographic and technological barriers when attempting to build annotated NLP datasets. We believe learners with high motivations and learning effects are more likely to be engaged in data annotation.

Limitations

This paper covers only four NLP tasks. Certain other tasks requiring more background knowledge may show different results. We suggest recruiting language learners when native speakers are not available, but recruiting learners may also be difficult for languages that are not popular for learners. Our results are based on a relatively low number of participants, as we chose to cover three different languages to show generalizability across languages. Many factors that may contribute to the results remain, such as the order of the batch of annotation questions with respect to the question difficulty level.

Ethics Statement

All studies in this research project were performed under KAIST Institutional Review Board (IRB) approval. We consider ethical issues in our experiments with language learners and native speakers.

The first consideration is fair wages. We estimated the average time per session (Step 2.1 to 2.3) based on a small pilot study and set the wage per session to be above the minimum wage in the Republic of Korea (KRW 9,160 $\approx$ USD 7.04) [17]. Table 9 shows that the expected hourly wages of all experiments exceed the minimum wage. We estimated the time for watching the orientation video and reading the instruction manual as one hour and provided compensation for this time of KRW 10,000.

There was no discrimination when recruiting and selecting the participants for the experiment, including all minority groups and factors such as age, ethnicity, disability, and gender. We used the sentences from publicly available datasets and manually excluded samples that may contain toxic and/or controversial contents.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics). This work was supported by a grant of the KAIST-KT joint research project through AI2XL Laboratory, Institute of convergence Technology, funded by KT [G01220613, Investigating the completion of tasks and enhancing UX]. Rifki Afina Putri was supported by Hyundai Motor Chung Mong-Koo Global Scholarship.

References

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Cash Costello, Shelby Anderson, Caitlyn Bishop, James Mayfield, and Paul McNamee. 2020. Dragonfly: Advances in non-speaker annotation for low resource languages. In *Proceedings of the 12th Language*

[17] https://www.minimumwage.go.kr/

Resources and Evaluation Conference, pages 6983–6987, Marseille, France. European Language Resources Association.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Ulf Hermjakob, Jonathan May, Michael Pust, and Kevin Knight. 2018. Translating a language you don't know in the Chinese room. In *Proceedings of ACL 2018, System Demonstrations*, pages 62–67, Melbourne, Australia. Association for Computational Linguistics.

Barbora Hladká, Jirka Hana, and Ivana Lukšová. 2014. Crowdsourcing in language classes can help natural language processing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2(1):71–72.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Siok H. Lee and James Muncie. 2006. From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, 40(2):295–320.

Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Chris Madge, Juntao Yu, Jon Chamberlain, Udo Kruschwitz, Silviu Paun, and Massimo Poesio. 2019. Progression in a language annotation game with a purpose. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):77–85.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243, Online. Association for Computational Linguistics.

Stephen Mayhew and Dan Roth. 2018. TALEN: Tool for annotation of low-resource ENtities. In *Proceedings of ACL 2018, System Demonstrations*, pages 80–86, Melbourne, Australia. Association for Computational Linguistics.

Lionel Nicolas, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. 2020. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of the 12th Language Resources and*

Evaluation Conference, pages 268–278, Marseille, France. European Language Resources Association.

Lucy Park. 2016. Naver sentiment movie corpus.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Federico Sangati, Stefano Merlo, and Giovanni Moretti. 2015. School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE*, pages 16–19.

Wenyuh Shieh and Mark R. Freiermuth. 2010. Using the dash method to measure reading comprehension. *TESOL Quarterly*, 44(1):110–128.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tatiana Tsygankova, Francesca Marini, Stephen Mayhew, and Dan Roth. 2021. Building low-resource NER models using non-speaker annotations. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 62–69, Online. Association for Computational Linguistics.

Luis von Ahn. 2013. Duolingo: Learn a language for free while helping to translate the web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, page 1–2.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Appendix

A Experiment Setup

A.1 Workflow

Post-survey Questions

- This task is difficult for me.
- I think my vocabulary skills have improved after doing this task.
- I think my grammar/structure skills have improved after doing this task.
- I consulted the additional resources often.
- Additional resources are helpful for completing the task.
- I am willing to participate in this task again.

A.2 Experiment Platform

All experiments were done on the website that we made and all responses and time taken are recorded. Figure 5 shows the screenshots of the pre-/post-test (a) and annotation (b) steps.

B Further Results

B.1 Annotation Quality

Languages and Tasks. Table 10 shows task difficulty with respect to four aspects: annotation accuracy, inter-annotator agreement, time, and perceived difficulty.

Majority Voted Labels. Table 11 shows the statistics of aggregated labels using majority vote. The number of splits means how many samples are not able to be aggregated in a single label (e.g., all annotators picked different labels, some annotators answered as *I don't know* so that it was too few to be aggregated), and the number of incorrect samples means how many samples are different to the ground truth labels.

Additional Resources. Table 12 shows whether the types of additional resources that learners consult affect annotation accuracy among three languages. English learners with the translation setting showed slightly better performance than those with the dictionary setting, while vice versa in Korean and Indonesian. It implies that it would be better to provide translations in high-resource languages with reliable machine translation systems, while mid- to low-resource language learners should consult dictionaries.

	EN	KO	ID
SA	$7.87_{\pm 0.30}$	$7.59_{\pm 0.21}$	$8.35_{\pm 0.16}$
NLI	$6.81_{\pm 0.20}$	$6.54_{\pm 0.18}$	$6.46_{\pm 0.26}$
NER	$8.59_{\pm 0.10}$	$8.44_{\pm 0.12}$	$7.12_{\pm 0.31}$
MRC	$7.11_{\pm 0.28}$	$7.18_{\pm 0.19}$	$7.95_{\pm 0.11}$

(a) Annotation accuracy

	EN	KO	ID
SA	$0.60_{\pm 0.03}$	$0.62_{\pm 0.02}$	$0.71_{\pm 0.03}$
NLI	$0.84_{\pm 0.01}$	$0.85_{\pm 0.01}$	$0.56_{\pm 0.01}$
NER	$0.76_{\pm 0.03}$	$0.76_{\pm 0.03}$	$0.85_{\pm 0.03}$
MRC	$0.63_{\pm 0.03}$	$0.51_{\pm 0.03}$	$0.53_{\pm 0.04}$

(b) Inter-annotator agreement measured by pairwise F1

	EN	KO	ID
SA	$3.34_{\pm 0.34}$	$3.28_{\pm 0.25}$	$2.43_{\pm 0.18}$
NLI	$4.48_{\pm 0.74}$	$9.72_{\pm 2.03}$	$8.50_{\pm 1.33}$
NER	$4.50_{\pm 0.38}$	$8.61_{\pm 0.88}$	$8.82_{\pm 1.01}$
MRC	$6.29_{\pm 0.57}$	$9.58_{\pm 0.72}$	$7.27_{\pm 0.41}$

(c) Time spent (minutes)

	EN	KO	ID
SA	$2.33_{\pm 0.88}$	$2.83_{\pm 0.31}$	$2.33_{\pm 0.42}$
NLI	$2.60_{\pm 0.60}$	$3.17_{\pm 0.48}$	$4.00_{\pm 1.00}$
NER	$3.40_{\pm 0.60}$	$3.17_{\pm 0.48}$	$3.50_{\pm 0.50}$
MRC	$1.67_{\pm 0.33}$	$3.83_{\pm 0.31}$	$3.00_{\pm 0.52}$

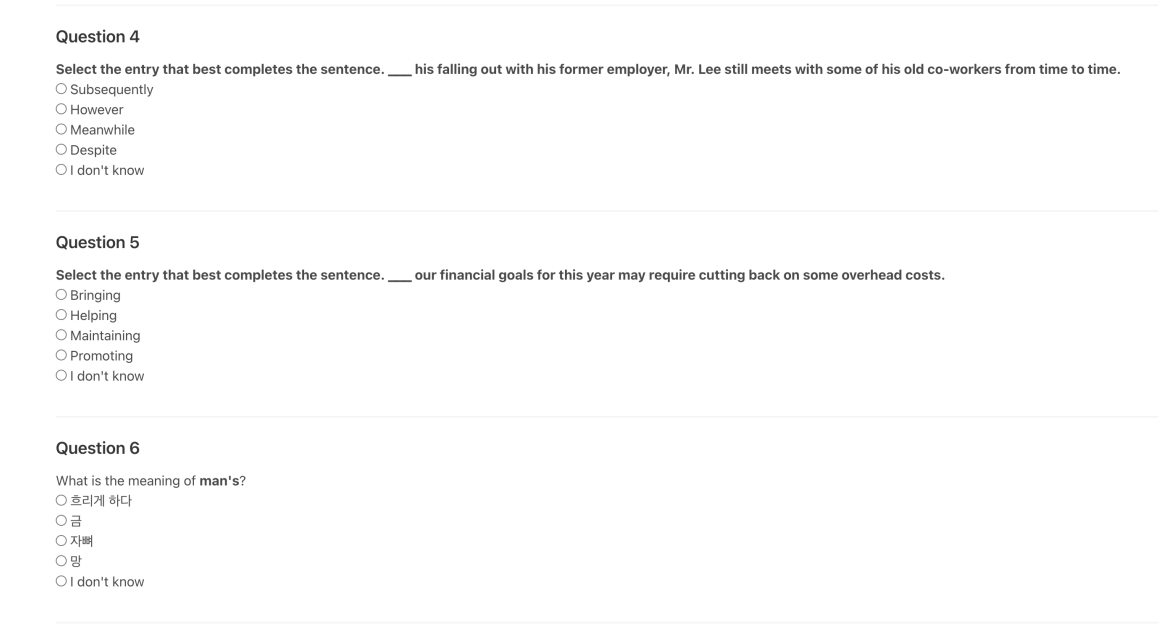
(d) Perceived difficulty from 1 *(very easy)* to 5 *(very hard)*

Table 10: Difficulty according to language and task

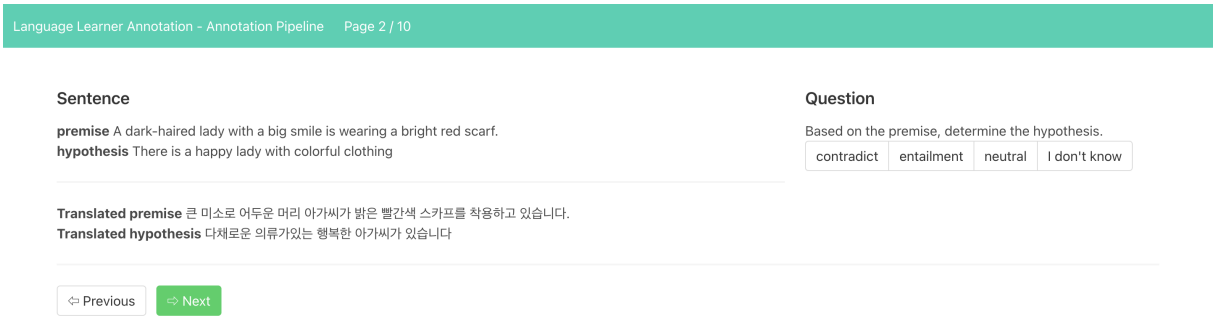
Language Proficiency and Native Speakers. We recruited three native speakers of each language and asked them to do the same experiments (pre-test, annotation, and post-test).

Table 13 shows the number of correct questions out of 10 and the time duration by each level of language learners and native speakers. Native speakers achieved the highest accuracy across all tasks taking the shortest time. It implies that there are some questions that native speakers can solve but learners cannot. We discuss those samples in Section 5.1. Time duration shows a significant gap between learners and native speakers, especially on NLI, but the gap was minimized at NER whose task requires annotators to tag all sequences.

B.2 Training Simulation with Learners' Annotation



(a) Pre-/post-test



(b) Annotation

Figure 5: Screen shot of experiment platform

Soft-labeled Synthetic Data. We tried training simulations with BERT-based models on synthetic data generated using soft labeling. We used soft labeling instead of majority voting to consider the variance among the annotators. Table 14 shows experimental results of BERT-based models on synthetic data whose data distributions come from the soft-labeled aggregations. It delivers similar findings to Table 5, while showing some noises. Models trained on native speakers' synthetic labels sometimes achieved similar performance to the Ground Truth while sometimes achieving the poorest performance such as EN-SA, EN-NLI, and KO-NLI. Our native annotators showed low inter-annotator agreement in those languages and tasks, so the synthetic labels based on native speakers' annotations were noisy.

		SA			NLI		
		EN	KO	ID	EN	KO	ID
Native Speaker	-	0 / 12	0 / 12	0 / 7	7 / 35	11 / 38	1 / 10
Language Learner	All	0 / 16	0 / 20	0 / 13	0 / 31	0 / 21	0 / 28
	Dictionary	0 / 23	0 / 16	0 / 13	4 / 39	2 / 27	0 / 31
	Translation	0 / 14	0 / 26	0 / 14	3 / 29	11 / 39	11 / 45
		NER			**MRC**		
		EN	KO	ID	EN	KO	ID
Native Speaker	-	3 / 21	1 / 19	2 / 18	4 / 23	6 / 19	3 / 20
Language Learner	All	4 / 17	6 / 21	3 / 20	5 / 22	7 / 24	6 / 25
	Dictionary	2 / 21	3 / 19	4 / 14	3 / 18	6 / 20	4 / 25
	Translation	5 / 19	3 / 15	2 / 17	6 / 27	4 / 14	5 / 16

Table 11: Majority vote results. (the number of splits / the number of incorrect) samples out of 120.

	Dictionary	**Translation**
EN	$7.21_{\pm 0.24}$	$7.84_{\pm 0.12}$
KO	$7.77_{\pm 0.12}$	$7.11_{\pm 0.15}$
ID	$8.14_{\pm 0.10}$	$7.05_{\pm 0.17}$

Table 12: Annotation accuracy with respect to languages and additional resources

Few-shot Learning using mT5. We also tried few-shot learning with mT5_{BASE} (Xue et al., 2021), a large-scale multilingual pretrained model which covers 101 languages including our target languages: English, Korean, and Indonesian. Table 15 shows that all models achieved comparable results to the baseline model within the margin of error. The gap among all models was relieved and we suppose that large-scale LMs with massive training data, including mT5, can perform too well on our common NLP tasks and our labeled data were too small to affect those models.

C Further Discussions

C.1 Learning Effect

Additional Resources. Table 19 (b) shows that both additional resources helped learners to remind or learn vocabulary used in the annotation samples.

Perceived Learning Effect. Table 17 shows similar trends to the previous results that basic-level learners perceived more learning effects on both vocabulary and grammar. They tend to show more willingness to re-participate in data annotation. Advanced-level learners show a high willingness to re-participate in data annotation, and this is because it was hard to improve their language proficiency. However, the sentences in data annotation were easy enough for them.

Table 18 shows self-rated language proficiency before and after the experiments when the description of CEFR criteria was given. Basic-level learners felt that their language proficiency had improved, while other levels of learners did not show a significant difference. Advanced-level learners tend to underestimate their language proficiency humbly.

Language Proficiency and Additional Resources. Table 19 (a) shows annotation accuracy compared to the ground truth labels concerning the learners' language proficiency level and the additional resources they used. There was no significant difference between the two settings with the learners either in the intermediate or the advanced level, while basic level learners achieved higher accuracy in dictionary settings. We suppose that basic-level learners might not be able to fill the gap of the wrong spans in the machine-translated sentence.

Table 19 (b)-(c) show users' responses on how frequently they consult additional resources and how helpful they were in data annotation. The frequency that the learners consult the additional resources and how the additional resources are helpful go together. All levels of learners replied that the dictionary setting was more helpful than the translation setting. Most basic-level learners in all

	Language Learners			Native Speakers
	Basic	Intermediate	Advanced	
SA	$6.96_{\pm 0.25}$	$8.26_{\pm 0.16}$	$8.43_{\pm 0.16}$	$9.00_{\pm 0.14}$
NLI	$6.96_{\pm 0.18}$	$6.38_{\pm 0.21}$	$6.64_{\pm 0.21}$	$7.59_{\pm 0.30}$
NER	$7.98_{\pm 0.12}$	$8.48_{\pm 0.09}$	$7.75_{\pm 0.28}$	$8.99_{\pm 0.07}$
MRC	$6.56_{\pm 0.26}$	$7.88_{\pm 0.12}$	$7.71_{\pm 0.13}$	$8.51_{\pm 0.14}$

(a) Annotation accuracy

	Language Learners			Native Speakers
	Basic	Intermediate	Advanced	
SA	$2.84_{\pm 0.24}$	$3.74_{\pm 0.26}$	$2.36_{\pm 0.18}$	$0.92_{\pm 0.07}$
NLI	$8.58_{\pm 0.69}$	$5.78_{\pm 1.23}$	$10.92_{\pm 3.46}$	$3.38_{\pm 0.59}$
NER	$7.95_{\pm 0.51}$	$5.83_{\pm 0.50}$	$7.68_{\pm 1.12}$	$7.23_{\pm 2.90}$
MRC	$7.03_{\pm 0.51}$	$9.53_{\pm 0.67}$	$7.13_{\pm 0.58}$	$4.85_{\pm 1.10}$

(b) Time duration

Table 13: Task difficulty between three levels of learners and native speakers with respect to annotation accuracy and time duration

		SA			NLI		
		EN	KO	ID	EN	KO	ID
Ground Truth	-	$89.56_{\pm 1.11}$	$85.29_{\pm 0.79}$	$97.20_{\pm 0.86}$	$79.05_{\pm 1.44}$	$79.00_{\pm 2.48}$	$68.20_{\pm 1.32}$
MT Dataset	-	$79.25_{\pm 1.25}$	$75.27_{\pm 1.33}$	$87.19_{\pm 1.39}$	$56.78_{\pm 2.26}$	$47.06_{\pm 1.26}$	$52.35_{\pm 1.35}$
Native Speakers	-	$70.66_{\pm 1.60}$	$84.66_{\pm 1.40}$	$96.48_{\pm 0.76}$	$67.67_{\pm 1.60}$	$56.18_{\pm 1.55}$	$67.12_{\pm 1.61}$
Language Learners	All	$85.75_{\pm 2.21}$	$80.22_{\pm 0.92}$	$92.37_{\pm 1.45}$	$78.38_{\pm 2.05}$	$72.51_{\pm 1.22}$	$61.99_{\pm 3.18}$
	Dictionary	$77.35_{\pm 1.92}$	$82.94_{\pm 1.09}$	$91.04_{\pm 0.65}$	$62.40_{\pm 2.86}$	$70.27_{\pm 1.89}$	$63.33_{\pm 2.24}$
	Translation	$85.29_{\pm 1.28}$	$72.98_{\pm 1.74}$	$90.40_{\pm 1.07}$	$68.88_{\pm 1.68}$	$65.61_{\pm 1.06}$	$56.54_{\pm 3.86}$

Table 14: Experimental results of BERT-based models trained on labels generated or synthesized by each group using soft-labeling

		SA			NLI		
		EN	KO	ID	EN	KO	ID
Ground Truth	-	$89.07_{\pm 3.45}$	$89.19_{\pm 2.85}$	$94.07_{\pm 3.62}$	$78.34_{\pm 2.47}$	$80.64_{\pm 3.20}$	$68.34_{\pm 2.82}$
MT Dataset	-	$85.13_{\pm 2.12}$	$84.57_{\pm 3.18}$	$90.10_{\pm 2.53}$	$74.48_{\pm 3.79}$	$77.92_{\pm 2.29}$	$63.20_{\pm 2.95}$
Native Speakers	-	$88.64_{\pm 3.11}$	$88.67_{\pm 3.54}$	$93.64_{\pm 2.12}$	$77.45_{\pm 2.85}$	$79.36_{\pm 3.13}$	$66.45_{\pm 3.58}$
Language Learners	All	$87.26_{\pm 3.10}$	$87.32_{\pm 2.56}$	$93.26_{\pm 3.13}$	$78.41_{\pm 3.46}$	$80.82_{\pm 2.70}$	$68.41_{\pm 3.15}$
	Dictionary	$88.16_{\pm 3.55}$	$88.28_{\pm 3.53}$	$94.16_{\pm 2.13}$	$76.64_{\pm 3.39}$	$81.08_{\pm 3.75}$	$69.64_{\pm 2.76}$
	Translation	$85.39_{\pm 2.71}$	$87.47_{\pm 2.34}$	$92.47_{\pm 2.88}$	$74.69_{\pm 2.99}$	$73.03_{\pm 2.65}$	$69.84_{\pm 3.19}$

Table 15: Experimental results of Few-shot Learning using mT5

languages consult and rely on additional resources.

There was no significant trend in the learners' frequency of consulting the additional resources concerning language and types of additional resources. Still, learners of all languages replied that the dictionary setting was more helpful for data annotation than the translation setting.

C.2 Feedback from Participants

Table 10 (c) shows perceived difficulty based on users' responses on post-survey. Participants responded that NER was the most complicated task

	Dictionary	Translation
pre-test	$3.63_{\pm 0.07}$	$3.32_{\pm 0.07}$
post-test	$3.53_{\pm 0.07}$	$3.41_{\pm 0.06}$

(a) Number of correct standardized test questions out of 5

	Dictionary	Translation
pre-test	$8.81_{\pm 0.08}$	$8.47_{\pm 0.08}$
post-test	$9.29_{\pm 0.07}$	$8.92_{\pm 0.07}$

(b) Number of correct word meaning questions out of 10

Table 16: Effect of additional resources in language learning with respect to language proficiency

	Basic	Intermediate	Advanced
vocab	$4.21_{\pm 0.13}$	$3.41_{\pm 0.13}$	$3.40_{\pm 0.13}$
grammar	$3.36_{\pm 0.13}$	$2.77_{\pm 0.13}$	$2.65_{\pm 0.13}$
willingness	$3.93_{\pm 0.21}$	$2.95_{\pm 0.21}$	$3.30_{\pm 0.21}$

Table 17: Users' responses on post-survey in terms of learning effect on vocabulary and grammar and willingness to re-participate

	Basic	Intermediate	Advanced
pre-survey	$0.57_{\pm 0.14}$	$2.45_{\pm 0.17}$	$3.20_{\pm 0.14}$
post-survey	$1.29_{\pm 0.19}$	$2.55_{\pm 0.14}$	$3.20_{\pm 0.17}$

Table 18: Self-rated language proficiency before and after data annotation experiment

and SA was the easiest. This result looks awkward considering that language learners achieved the highest accuracy in NER.

Learners replied that exactly distinguishing the start and the end of the named entity was confused in NER, and some named entities were unfamiliar with them if they were not used to the domain. All learners in the translation-provided setting on NLI replied that the machine-translated sentences were incorrect and even disturbing to infer the textual entailment between two sentences. Most Indonesian learners on SA replied that the sentences usually contain multiple sentiments, representing that some points are good, but others are bad, so they are unsure about their labels. This is probably due to the characteristics of IndoLEM (Koto et al., 2020) whose sentences come from Hotel reviews with multiple features. Learners should read a passage in MRC so that it helps to improve their language proficiency, while advanced-level learners who are fluent in the target language replied that they do

	Dictionary	Translation
Basic	$7.40_{\pm 0.18}$	$6.90_{\pm 0.17}$
Intermediate	$7.86_{\pm 0.12}$	$7.60_{\pm 0.13}$
Advanced	$7.99_{\pm 0.14}$	$7.31_{\pm 0.16}$

(a) Annotation accuracy

	Dictionary	Translation
Basic	$4.67_{\pm 0.21}$	$3.25_{\pm 0.59}$
Intermediate	$2.75_{\pm 0.45}$	$2.70_{\pm 0.42}$
Advanced	$2.75_{\pm 0.41}$	$2.75_{\pm 0.39}$

(b) Frequency of consulting additional resources

	Dictionary	Translation
Basic	$4.67_{\pm 0.21}$	$3.75_{\pm 0.53}$
Intermediate	$3.42_{\pm 0.47}$	$3.30_{\pm 0.37}$
Advanced	$3.62_{\pm 0.46}$	$3.25_{\pm 0.41}$

(c) Help of additional resources

Table 19: Effect of additional resources with respect to language proficiency

not have to read the whole passage but read the sentence that contains the answer span.

D Qualitative Analysis

D.1 Failure Reason Analysis on Learners' Annotation

Table 20 shows the examples of three failure reasons: ungrammatical sentence, task ambiguity, and culturally-nuanced expression. Missing period between two short sentences in the SA sample (a) leads to misunderstandings among learners. Ambiguity, whether "*people*" and "*some people*" in premise and hypothesis are indicating the same in (b), leads all learners and native speakers to get confused between `neutral` and `contradiction`, which is an ambiguity of NLI itself. "*ajaran yang dipercayai*" in questions in the MRC sample (c) literally means "*teachings believed by*" in Indonesian, but its correct translation is "*belief*" or "*religion*". Learners failed to interpret those difficult and culturally-nuanced expressions correctly and generated wrong labels, while all native speakers found the same answer.

D.2 Qualitative Analysis on Pre-/Post-test

We analyze the characteristics of pre- and post-test questions that the learners get wrong. For English, two questions that every learner got wrong were

GRE questions, which are notably difficult even for native speakers. Many learners picked the "*I don't know*" option for GRE questions as well. For Korean, there was no question that every learner got wrong. However, for A-level learners, a large number of them answered 'Arrange the sentences in the correct order' questions incorrectly. The difficulty may stem from their insufficient knowledge of transition signals and the logical flow in the target language. Also, learners chose "*I don't know*" option a lot for questions requiring an understanding of newspaper titles. For Indonesian, learners mostly fail on questions related to prepositions, prefixes and suffixes, and formal word formation.

Most of the questions that most learners answered incorrectly require an understanding of the context and the grammatical structure. These aspects of language are difficult to learn within a short time, attributing to the insignificant difference in the scores between the pre- and post-tests.

	Lang.	Level	Type	Sentence	Ground Truth	Language Learners	Native Speakers	Failure Reason
(a)	KO	1	Original	스토리가 어려움 볼만함	pos	neg	pos	Ungrammatical sentence
			Machine Trans	Story is difficult to see				
			Correct Trans	The story if difficult, [but it's] worth watching.				
(b)	EN	2	Original	[Premise] People standing at street corner in France. [Hypothesis] Some people are taking a tour of the factory.	con	neu	neu	Task ambiguity
			Machine Trans	[Premise] 프랑스의 거리 모퉁이에 서있는 사람들. [Hypothesis] 어떤 사람들은 공장을 여행하고 있습니다.				
			Correct Trans	[Premise] 프랑스의 거리 모퉁이에 서있는 사람들. [Hypothesis] 어떤 사람들은 공장을 관광하고 있습니다.				
(c)	ID	5	Original	[Context] ... Mereka membangun komunitas dengan berpegang teguh pada spiritualitas sebagai dasar pembentukan ajarannya. Tidak jarang pula mereka menyebut kepercayaannya sebagai agama Jawa. Melalui kepercayaan ini, mereka melakukan penggalian kembali kepercayaan dan nilai-nilai spiritualitas masyarakat Jawa masa lalu, terutama pada masa prapatrimonial. ... [Question] Apakah ajaran yang dipercayai Suku Dayak Hindu Budha Bumi Segandu Indramayu?	agama Jawa	*I don't know*; spiritualitas; etc	agama Jawa	Culturally-nuanced expression
			Machine Trans	[Context] ... They built the community by clinging to spirituality as the basis for the formation of his teachings. Not infrequently also they mentioned his belief as Java religion.Through this belief, they re-excavated trust and the spirituality values of the past Javanese society, especially during the praprimonial period. ... [Question] What is the teachings believed by the Hindu Buddhist Bumi Division of Indramayu?				
			Correct Trans	[Context] ... They built the community by clinging to spirituality as the foundation for their teachings formation. Not infrequently, they called their belief as Java religion. Through this belief, they re-dig trust and the spiritual values of the past Javanese society, especially during the pre-patrimonial period. ... [Question] What is the teachings believed by *(belief/religion)* the Dayak Hindu Buddha Bumi Segandu Indramayu?				

Table 20: Example annotation questions that all learners fail

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section <Limitations>

☑ A2. Did you discuss any potential risks of your work?
Section <Limitations>

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section <1. Introduction>

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section <3. Study Design>, <4. Experimental Results>, <5. Discussion>

☑ B1. Did you cite the creators of artifacts you used?
Section <References>

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We only used scientific artifacts from research papers that are publicly available.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section <3. Study Design>, <Appendix>

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section <3. Study Design>, <4. Experimental Results>, <5. Discussion> <Appendix>

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section <3. Study Design>, <4. Experimental Results>, <5. Discussion> <Appendix>

C ☑ Did you run computational experiments?

<Section 4.2. Training Simulation with Learners' Annotations>

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
<Section 4.2. Training Simulation with Learners' Annotations>

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3 <Study Design>

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section <Appendix>

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 9 <Ethics Statement>

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 9 <Ethics Statement>

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 9 <Ethics Statement>

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3 <Study Design>