Developing a Typology for Language Learning Feedback

Steven Coyne[1,2] Diana Galvan-Sosa[1,2] Keisuke Sakaguchi[1,2] Kentaro Inui[1,2]
[1]Tohoku University [2]RIKEN
coyne.steven.charles.q2@dc.tohoku.ac.jp
{dianags,keisuke.sakaguchi,kentaro.inui}@tohoku.ac.jp

Abstract

Writing is an important part of language learning. With the recent release of corpora containing feedback on learner writing, it has become easier for NLP researchers to examine this process and work towards such tasks as automatic feedback comment generation. However, analysis and generation are hindered by a lack of a typology for such comments, and it is costly to determine frequency distributions or generation error rates for different kinds of comments. In this paper, we discuss typologies from both NLP and educational research, and propose a system to combine them to create an annotation scheme for feedback comments.

1 Introduction

Written corrective feedback on learner text is widespread in language education, and an active area of research in the field of second language acquisition [1]. Research has shown that properly administered written feedback has a positive effect on language learning [2, 3], including in electronic settings [4]. Analysis of the content of these feedback comments, as well as where and when they are added by instructors, is of great interest to educators, educational technology developers, and NLP researchers focusing on learner writing.

In addition, it is ideal if realistic comments can be generated automatically during assessment in learning environments. This would save teachers time and energy by providing suggestions and allowing them to accept or edit suitable ones while rejecting unsuitable ones. Similar technology can be used to provide feedback comments directly to learners in an intelligent tutoring setting.

In NLP, work on language education feedback comments has mostly been from the perspective of feedback comment generation, defined as the task of generating hints or explanatory notes for language learners [5]. Data consists of learner sentences, associated feedback comments, and offsets highlighting where the comments were attached to the sentence. There are very few corpora of such sentence pairs, and work on them has mostly focused on the ICNALE Learner Essays with Feedback Comments Dataset[6],[1] summarized in in Table 1. There is also a translated subset of this dataset used in GenChal 2022 [7], and a separate corpus focusing on transition word comments [8].

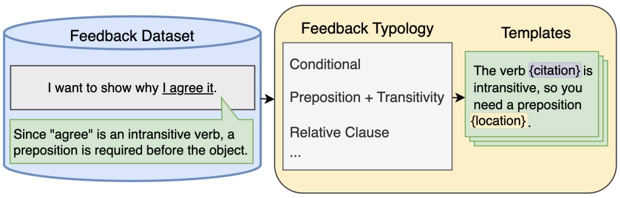

Feedback Dataset
I want to show why I agree it.
Since "agree" is an intransitive verb, a preposition is required before the object.
Feedback Typology
Conditional
Preposition + Transitivity
Relative Clause
...
Templates
The verb {citation} is intransitive, so you need a preposition {location}.

Figure 1 Visualization of the use of manually-labeled feedback comments to support the development of a template-based feedback comment generation system.

One challenge to working with feedback comment data is the lack of a dedicated typology for feedback comments and their connection to the original sentence. If such a typology existed, it could enable more detailed analysis of these corpora and support certain approaches to feedback comment generation, as detailed in Figure 1.

Educational research characterizing feedback comments often focuses on the learning effect of comments based on dimensions such as directness [2], presence of metalinguistic terms [3], and hedging [9]. However, these approaches do not touch on the comment's editorial intent. Meanwhile, the closest typologies from NLP are the error types seen in the field of Grammatical Error Correction (GEC). These error types model grammatical errors well, but can not capture many broader topics that instructors tend to comment on, as described in section 2. We seek to develop a typology specifically for the context of feedback comment corpora consisting of learner sentences and instructor comments.

[1] The dataset is available at https://www.gsk.or.jp/en/catalog/gsk2019-b

General	
Sentences	43568
Feedback Comments	26592
Commented Sentences	19991
Commented Sentence Ratio	0.459
Most Comments on One Sentence	14

Preposition	
Sentences	28829
Feedback Comments	5693
Commented Sentences	4931
Commented Sentence Ratio	0.171
Most Comments on One Sentence	6

Table 1 Information about the "ICNALE Learner Essays with Feedback Comments" dataset. It is divided into two sub-corpora, one with comments on general topics, and the other focusing on preposition use.

2 Background

The type of feedback comment data we discuss in this work was first presented in the context of feedback comment generation [5]. Challenges revealed in subsequent work in this subfield serve as motivation for creating a dedicated typology. Several issues were highlighted by Hanawa et al. [10] and the participants of GenChal 2022. First, generation is confounded by a many-to-one problem in which multiple comments can have the same topic:

(1)***We reached to the station.**

Because the verb "reach" is a transitive verb, the preposition "to" is not required.

(2)***I reached to New York.**

"Reach" is a transitive verb. This verb does not require a preposition prior to the object.

These comments are about the same error, but the text content is superficially different. This can result in mixed generations which are less clear [10].

Furthermore, there are a large number of very specific comments relating to particular words and their collocations. The number and diversity of such comments contributes to the mixed output problem as well.

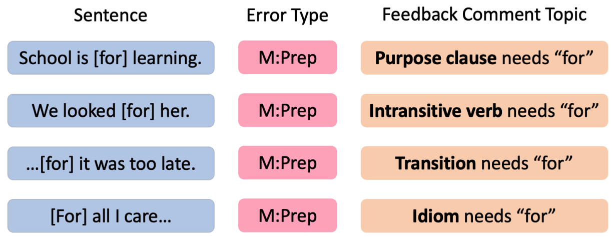
Finally, generation can result in inaccurate or misleading comments [10]. It is important to constrain these false generations, which can have a negative learning effects or reduce confidence in the system.

It is difficult to address the above issues without first identifying in greater detail what kinds of feedback comments exist and how these issues manifest across those categories. It is possible that a few kinds of feedback comments are responsible for a large portion of generation errors. Currently, such investigation would involve manual analysis using ad-hoc categories, and any further analysis by another team would involve a similar process starting over from scratch. It is therefore beneficial to develop a categorization system for data from this task.

When approaching the issue of how to tag feedback comments on learner sentences, a natural first step is to consider the error type tags which have been used in the NLP subfield of GEC. These include the system used in the NUCLE dataset [11], the system used in Cambridge Learner Corpus [12] and seen in the First Certificate in English (FCE) dataset [13], and the system from ERRANT [14], which has become the dominant system in GEC research at this time. These typologies model grammatical errors fairly well, often by using a combination of parts of speech and the actions of deletion, insertion, or replacement to describe the errors.

There are many cases where an instructor's comment identifies a grammatical error in a straightforward manner easily modeled by these systems. However, there are many additional cases in which the highly local and orthographically-focused GEC categories can not fully characterize the issue identified in the comment. The feedback comments instructors write often refer to complex errors or combinations of errors. Even a simpler operation such as adding the preposition "for" could be done for a number of reasons, such as establishing a purpose clause, completing an idiom, or to serve as a transition word (see Figure 2). Meanwhile, the ERRANT system would tag all such cases as "M:PREP" (Missing Preposition). This describes the desired edit, but not the reason for the edit.

The reverse is also true, as any number of errors could be connected to a comment about an "unclear" sentence. These many-to-one and one-to-many cases suggest the need for additional tags concerning "higher level" concepts common in instructor comments, such as purpose clauses and conditionals. Furthermore, some common types of feedback comments are broader than grammatical error correction itself. There are comments about idiom, tone, the argument made by a sentence, and praise, resulting in feedback which may not align with GEC error types at all.



Sentence	Error Type	Feedback Comment Topic
School is [for] learning.	M:Prep	**Purpose clause** needs "for"
We looked [for] her.	M:Prep	**Intransitive verb** needs "for"
...[for] it was too late.	M:Prep	**Transition** needs "for"
[For] all I care...	M:Prep	**Idiom** needs "for"

Figure 2 Abstraction of a one-to-many relationship between an ERRANT error type and different kinds of feedback comments.

3 Method

To address this, we can search beyond NLP error typologies and look to the field of education. One direct source of topics for teacher comments is the various sets of error code annotations used by English teachers. These tend to include many of the higher-level categories we wish to address, such as redundancy and idiom. While there does not seem to be a well-accepted correction code standard in literature, there are a variety of systems shared online, many covering similar topics. One example is the system used for writing programs at the University of California, Irvine [15]. Ideally, a classification system for instructor feedback comments can incorporate some of these principles when the GEC error types can not fully express the content of the comment.

We propose a loosely hierarchical tagging system in which the "lower" level tags correspond to GEC-style error types such as "Unnecessary Preposition," and "higher" level tags are used to characterize comments which exceed the scope of GEC, such as praise, idiom corrections, and advice about forming specialized clauses such as conditionals and dummy subject clauses. For a given sentence-comment pair, we apply the highest level tag which can accurately characterize the comment's topic.

The tags are developed by first consulting previous typologies, considering which categories are most likely to be used by English language teachers.[2] Furthermore, we define several principles for the development of the tag set:

1. It exists to descriptively model a phenomenon in natural language, the content of feedback comments. Anything common in this domain should be incorporated.
2. For now, we focus on the the sentence level, not considering the broader coherence of ideas in the text.
3. The categories should be easily understandable to educators and potential corpus annotators.
4. Each feedback comment should be labeled with a single tag which best characterizes it.
5. For comments about issues in the text, the relevant grammar point or type of clause is identified.
6. Tags are applied in hierarchical order, except when a comment clearly focuses on the lower level topic.

After drafting candidate tags, we test them against the contents of the ICNALE Learner Essays with Feedback Comments dataset, adapting to the realities of the corpus sentences. As a first step, we considered a subset consisting of 500 sentences, 250 each from the General and Preposition sub-corpora. We consider only sentences with exactly one feedback comment, and comments extending across multiple sentences were excluded, since they exceed the sentence-level scope of this work. Sentences were then sampled with a particular random seed.

The current high-level tags are given in Table 2, and the remainder of the current tag set is presented in the appendix. Work is ongoing, and the proposed tag set may evolve further by the time all sentences in the dataset have been considered and annotated.

4 Discussion

As the tag set developed, we obtained several insights. The frequency of certain topics was not always in accordance with expectations. Praise comments were very common, necessitating a dedicated tag for them. Comments proposing complex, specific rewrites were also quite common, leading to a "Rewrite" tag for comments of this nature. Depending on goals of developers or researchers, it may be desirable to exclude such comments from generation. It is thus prudent to tag them at this stage to allow toggling, and to facilitate any future attempts to classify comments as praise or direct rewrite suggestions.

It was also necessary to determine how fine-grained to make the tagging system. Ideally, there is enough coverage to describe the majority of instructor comments in a meaningful way, with only a few placed into an "other" category. The number of tags should be expressive enough to allow for disambiguation of cases as seen in Figure 2. At the same time, a large number of rare categories may not be as interpretable, and may be more difficult for annotators. We will continue to consider this balance as work progresses.

[2] The first author worked in English education for five years, and drew on that experience in the process.

High-Level Tags	
Tag Name	**Example**
Comparative	Maybe you will study [**more hard** → **harder**] in the class.
Causative	It will ruin our concentration and make everything [**getting**] worse.
Conditional	If I [**have** → **had**] a job, I could buy more things.
Dummy Subject	It is important [**that**] university students [**have**] a part time job.
Fragment	**Obligation at home and at campus.**
Idiom	[**There's** → **That's**] the way it goes.
Modal/Auxiliary	Students [**would better** → **should**] have part-time jobs.
Parallel Structure	...hanging out with my best friend, [**buy** → **buying**] cosmetics, or shopping
Praise	(Various kinds of praise and encouragement)
Redundancy	I did part-time jobs last summer vacation to [**go travel**] to a foreign land.
Relative Clause	College students [**who**] jump in part-time job have a variety of reasons.
Rewrite	(Used for explicit, complex revision suggestions)
Tone	It's maybe [**cause** → **because**] my work experience less than other people.
Unclear	**If home is not richness economically, everybody is only just doing it.**

Table 2 Hierarchical Annotation System for Feedback Comment Topics, High-level tags

5 Future Work

With categorized feedback comments, it becomes possible to compare the comments from each category against other information. One such possibility is to cluster the sentences and compare the clusters to the manual labels. This can be used to judge the variability of different kinds of feedback comments. [16] attempted to address the superficial diversity issue by clustering the comments with textual similarity, but the interpretability of the resulting categories is limited. It would be useful to compare such results to class labels added by a human.

In addition to surface similarity, we will experiment with clustering based on semantic similarity or with a topic-modeling approach as seen in [17]. Topic labels can be identified in the feedback comment text, and placed into hierarchical clusters. Given that there are many synonyms for grammatical terms in the feedback comments, we hypothesize that semantic or topic modeling will perform better than surface similarity. We will compare clusters to the manual tags, potentially revealing additional topic subtypes which can help improve the tagging logic.

It is also possible to compare the manual tags to the distribution of tags from various NLP tools. These include sequence taggers for parts of speech and dependencies as well as error correction systems. If some feedback comment topics correlate very strongly with certain parse patterns or GEC error types, those system outputs may be useful as predictive features for the feedback comments.

Furthermore, it is possible to replace the outputs of highly diverse comment categories with generalized templates. Manual tags allow us to identify the categories which most need such attention. The data may simplify if such comments are unified into a limited number of semi-automatically generated templates with slot-filling. The slots can be filled with words from the original sentence and information from lexical resources. This can help with the reliability challenge in feedback comment generation, since we can more tightly control outputs in these cases.

Creating templates also allows us a chance to rewrite comments to be more suitable to the task. We find that many of the comments in the ICNALE feedback dataset have fairly advanced grammatical explanations, which can be simplified to help learners understand them. Furthermore, there are comments with an error diagnosis, comments with edit suggestions, and comments containing both. If these are differentiated in the template system, developers or users of a comment generation system can decide whether to disable outputs from one or the other.

6 Conclusion

To assist with the analysis of written feedback comments for language learning, we propose a system to annotate feedback comment data with comment topics combining principles from grammatical error correction and the field of education, applying it to a subset of the ICNALE Learner Essays with Feedback Comments dataset as a first step in a plan to annotate the entire corpus. In the context of feedback comment generation, we describe how these tags can be used to assist with the analysis of data and the creation of template-guided generation systems.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP22H00524 and JP21K21343.

References

[1] Eun Young Kang and ZhaoHong Han. The efficacy of written corrective feedback in improving l2 written accuracy: A meta-analysis. **The Modern Language Journal**, Vol. 99, pp. 1–18, 2015.

[2] Dana Ferris and Barrie Roberts. Error feedback in l2 writing classes: How explicit does it need to be? **Journal of Second Language Writing**, Vol. 10, No. 3, pp. 161–184, 2001.

[3] John Bitchener. Evidence in support of written corrective feedback. **Journal of Second Language Writing**, Vol. 17, No. 2, pp. 102–118, 2008.

[4] Estela Ene and Thomas A. Upton. Learner uptake of teacher electronic feedback in esl composition. **System**, Vol. 46, pp. 80–95, 2014.

[5] Ryo Nagata. Toward a task of feedback comment generation for writing learning. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3206–3215, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. Creating corpora for research in feedback comment generation. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 340–345, Marseille, France, May 2020. European Language Resources Association.

[7] Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. Shared task on feedback comment generation for language learners. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 320–324, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[8] Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. A dataset for investigating the impact of feedback on student revision outcome. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 332–339, Marseille, France, May 2020. European Language Resources Association.

[9] Wendy Baker and Rachel Hansen Bricker. The effects of direct and indirect speech acts on native english and esl speakers' perception of teacher written feedback. **System**, Vol. 38, No. 1, pp. 75–84, 2010.

[10] Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. Exploring methods for generating feedback comments for writing learning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 9719–9730, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 22–31, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[12] Diane Nicholls. The cambridge learner corpus: Error coding and analysis for lexicography and elt. **Proceedings of the Corpus Linguistics 2003 conference**, p. 572–581, 2003.

[13] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 180–189, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[14] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 793–805, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[15] UCI Writing Center. Correction symbols for uci writing programs, 2008.

[16] Ryo Nagata and Kazuaki Hanawa. Kasotekina ayamari taipu no wariate ni yoru kaisetsubun seisei no seino kojo [improving the performance of commentary generation by assigning virtual error types]. In **Proceedings of the 27th Annual Conference of the Association for Natural Language Processing**, pp. 679–684, March 2021.

[17] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

Mid-Level Tags

Tag Name	Example
Derivation	Due to the time, we lived in a [**peace** → **peaceful**] world.
Hyphenation	It is important for students to have a [**part time** → **part-time**] job.
Nominalization	This is for [**keep** → **keeping**] fresh air in the place.
Noun Countability	Also, they can buy other [**stuffs** → **stuff**].
Participle	In some restaurant, we can see students [**works** → **working**] as waiters.
Passive Voice	As a result, their performance in school may be [**get**] influenced.
Possessive	Studying is the main task [**to** → **of**] students.
Preposition + Transitivity	I completely agree [**with**] this opinion.
Purpose Clause	They should earn money [**for** → **to**] spend in the daily life by themselves.
Quantifier	Almost [**all**] non-smokers hate the cigarette smoke.
Question Formation	Why [**students must** → **must students**] do part time job[**.** → **?**]
Run-on Sentence	In a word, I'll try[**,** → **.**] if I find a job fit me, I'll do that!
Subject-Verb Agreement	The [**students works**] part time job
Transitions	[**But** → **However,**] it costs a lot to go to the university.
Word Order	**What more serious is...** → **What is more serious...**

Low-Level Tags

Tag Name	Example
Capitalization	In [**korea** → **Korea**], it is common.
Incorrect/Double Negative	If smoking [**not be** → **is not**] banned, a lot of people will smoke.
Missing Adjective	Almost [**all**] restaurant in Japan have smoking seat.
Missing Adverb	And [**when**] they can get right answer, I feel very happy.
Missing Article	They will relax after having [**a**] meal.
Missing Noun	For students who don't have money, [**jobs**] are very necessary.
Missing Preposition	70% [**of**] men in this country is smoking
Missing Pronoun	Try to tell them what [**they**] should do, and what [**they**] should not to do.
Missing Verb	Some of them can not [**pay**] their education fees.
Noun Number	College students have a lot of [**times** → **time**].
Other	(Miscellaneous Topics)
Punctuation	They can learn the value of money[**,**] they use, too.
Replace Adjective	It 's [**interested** → **interesting**] to me .
Replace Adverb	I have [**ever** → **never**] been in this situation.
Replace Article	Second, they can know [**an** → **the**] importance of money.
Replace Noun	I will talk about my [**opinion** → **reason**] why.
Replace Preposition	I have three reasons [**about** → **for**] it.
Replace Pronoun	They need work for them or [**they** → **their**] family .
Replace Verb	It [**does** → **is**] important and helpful when taking a job.
Spacing	Customers [**may be** → **maybe**] don't want to go that restaurant again.
Spelling	[**The** → **They**] will make good use of the money.
Unnecessary Adjective	And it will be very [**important**] worthwhile in life.
Unnecessary Adverb	I feel bored every time [**when**] someone smokes near me.
Unnecessary Article	Nowadays it is [**a**] common for college students to have a part-time job.
Unnecessary Noun	Students have burden on a lot of assignments and expensive tuition [**fee**].
Unnecessary Preposition	Many students had a part-time job because they need [**to**] money.
Unnecessary Pronoun	I have acquaintarces that [**he**] died from smoking.
Unnecessary Verb	Many of people [**are**] get a part time job for many reasons.
Verb Conjugation	Smoking [**are** → **is**] very popular these days.
Verb Form	How about [**give** → **giving**] sometime to think yourself.
Verb Tense	Most students [**are** → **were**] isolated from society before.

Table 3 Hierarchical Annotation System for Feedback Comment Topics, Mid-level and Low-level tags