Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 332–339
Marseille, 11–16 May 2020

A Dataset for Investigating the Impact of Feedback on Student Revision Outcome

Ildikó Pilán[1*], John Lee[2], Chak Yan Yeung[2], Jonathan Webster[2]
[1]Department of Informatics, University of Oslo
[2]Department of Linguistics and Translation, City University of Hong Kong
ildikop@ifi.uio.no, jsylee@cityu.edu.hk, cyyeung91@gmail.com, ctjjw@cityu.edu.hk

Abstract

We present an annotation scheme and a dataset of teacher feedback provided for texts written by non-native speakers of English. The dataset consists of student-written sentences in their original and revised versions with teacher feedback provided for the errors. Feedback appears both in the form of open-ended comments and error category tags. We focus on a specific error type, namely linking adverbial (e.g. *however*, *moreover*) errors. The dataset has been annotated for two aspects: (i) revision outcome establishing whether the re-written student sentence was correct and (ii) directness, indicating whether teachers provided explicitly the correction in their feedback. This dataset allows for studies around the characteristics of teacher feedback and how these influence students' revision outcome. We describe the data preparation process and we present initial statistical investigations regarding the effect of different feedback characteristics on revision outcome. These show that open-ended comments and mitigating expressions appear in a higher proportion of successful revisions than unsuccessful ones, while directness and metalinguistic terms have no effect. Given that the use of this type of data is relatively unexplored in natural language processing (NLP) applications, we also report some observations and challenges when working with feedback data.

Keywords: corrective feedback, hedging, error correction, Intelligent Computer-Assisted Language Learning

1. Introduction

Feedback has an important role in facilitating the learning process (Gass and Mackey, 2013). Information about the occurrence of an error, explanation about why it occurred and suggestions about how to correct it can help language learners revise incorrectly used linguistic elements and it can contribute to avoiding repeated future occurrences of these errors. Despite numerous investigations on the kinds of feedback that can best promote learning, this question remains an open debate in the area of Second Language Acquisition (SLA) (Lightbown and Spada, 1990; Lyster and Ranta, 1997; Sheen and Ellis, 2011). The application of computational linguistic methods to answer this question remains much less explored (Nagata, 2019). Since these methods allow for the automatic processing of larger amounts of texts, it has a good potential to contribute to this debate with empirical evidence.

To promote research on feedback in the above mentioned areas, we collect and annotate a set of teacher feedback and pairs of original and revised student-written sentences from a corpus of teacher-corrected academic essays authored by non-native speakers of English[1]. We annotate two types of information: (i) *feedback directness*, whether the feedback explicitly contained the correction (direct) or not (indirect), and (ii) *revision outcome*, i.e. whether an error was successfully revised by a student or not. The resource enables investigations about which characteristics of teacher feedback influence students' revision outcome most, addressing thus the current lack of computational linguistic resources for feedback generation and in-depth data-driven analysis. This line of research can help identify the kind of feedback that is most beneficial for students and should therefore be prioritized in both human-human and human-computer interactions. Concrete examples of NLP applications in this context include (i) automatic feedback generation providing detailed explanations about errors to students; (ii) teacher support in e-learning platforms giving information about the expected revision outcome for the type of feedback being provided.

Some currently available online learning or writing platforms do offer automatically generated feedback, but these are often only available for shorter student input and they consist mainly of explicit corrections or error categories, with only a limited number of cases including any explanation (Vawter and Martens, 2019; Rudzewitz et al., 2018). A deeper understanding about how feedback can be best delivered could significantly increase the successful adoption of e-learning systems.

As revising some types of errors may inherently be harder than others, we control for this factor by collecting a subset of data targeting a specific error type, namely *linking adverbial* errors. Linking adverbials, often also referred to as *connectors*, are single words (mainly adverbs) or phrases (e.g. prepositional phrases) that express a logical connection to another phrase or sentence such as addition, e.g. *moreover* (Liu, 2008). The choice of linking adverbials as error category was motivated by a number of aspects. Firstly, they are an integral part of second language (L2) English writing syllabuses and are a rather common source of error (Larsen-Walker, 2017). Secondly, similar studies on feedback for such discourse-level error types is lacking even in the SLA literature (Gass and Mackey, 2013, p. 27).

In the following, we first provide an overview of the related literature in Section 2 and present the source corpus in Sec-

*Work performed mainly while at City University of Hong Kong.

[1]This corpus and the annotated dataset described here are available for research purposes through arrangement with the Halliday Centre for Intelligent Applications of Language Studies at City University of Hong Kong (hcls@cityu.edu.hk).

tion 3. We then describe the data extraction criteria and their implementation, as well as the annotation scheme and process in Sections 4 and 5 respectively. In Section 6, we present a statistical analysis on the resulting dataset targeting the effect of feedback type and feedback characteristics on revision outcome. The characteristics we focus on include directness based on the manual annotations as well as the use of metalinguistic terms and hedging, i.e. employing mitigating expressions. We find that successful revisions are more common with feedback consisting of open-ended comments rather than with error category tags. Moreover, feedback with hedging appears more often with successful revisions than with unsuccessful ones. Directness and the presence of metalinguistic terms, however, show no significant effect on revision outcome. We conclude by discussing some considerations about the automatic processing of this type of data in Section 7, in an attempt to provide useful insights for others planning to conduct research on this kind of educational data.

2. Related Work

2.1. Feedback in SLA

Feedback in SLA has been extensively studied, where a substantial body of research examined its characteristics and how those influence its ability to promote learning (Gass and Mackey, 2013). Corrective (negative) feedback, pointing learners to errors made and often accompanied by explanations or the correct form, has been especially of interest. Corrective feedback can occur in different forms and a number of different taxonomies have been proposed in previous work to categorize it. On the one hand, Lyster and Ranta (1997) define six categories of oral feedback based on discourse features. On the other hand, Sheen and Ellis (2011) propose a psycho-linguistically motivated taxonomy for written corrective feedback based on two dimensions, which are presented in Table 1. An example of indirect non-metalinguistic feedback would be the following comment: *fourthly is not used usually*.

	Direct	**Indirect**
Metalinguistic	correct form with explanation	error codes, explanation
Non metalinguistic	correct form only (no explanation), reformulation	correct form not provided, errors may or may not be located

Table 1: Written corrective feedback taxonomy.

We opt for adopting the taxonomy in Table 1 as it targets written language and allows for a more straightforward adaptation to computational methods. Although, a clear general consensus has not been reached about what type of feedback promotes learning most, previous investigations found that direct and metalinguistic feedback are more effective for, among others, learning article use (Sheen, 2007).

A number of studies have concentrated on the effect of hedging in teacher feedback. *Hedges* are linguistic devices which express "tentativeness and possibility in communication" with the aim of mitigating what is being said (Hyland, 1998, p. 1). The two main categories of hedging are: (i) *lexical hedging*, the most common type, achieved via the use of specific words and phrases (e.g modal verbs such as *may could*); and (ii) *strategic hedging* (e.g. reference to limiting conditions or a theory). Hedged teacher comments have been found to take more time to recognize by learners and produce less accurate corrections compared to a more direct phrasing in the feedback (Baker and Bricker, 2010).

Besides its characteristics, a number of other factors also influence the efficiency of feedback, such as the timing for providing it, the type of error and learners' proficiency level (Gass and Mackey, 2013).

2.2. Feedback in NLP

There is increasing interest in automatically generating feedback comments to text written by language learners. Research on this task, known as "feedback comment generation", has been mostly focused on preposition errors, with feedback generated by case frames (Nagata et al., 2014) and neural retrieval-based method (Nagata, 2019).

Automatically generated feedback is more typical for exercises and short answers, e.g. in Duolingo[2] (Settles et al., 2018), while its provision for entire texts has been less commonly addressed. An effort in this direction is presented in Rudzewitz et al. (2018), who describe a feedback generation method for a variety of error types within a web-based workbook, *Feedbook*. The approach consists of the offline generation of potential ill-formed responses and an online matching mechanism to the actual student responses. Feedbook provides *scaffolding feedback*, i.e. indirect metalinguistic feedback that guides students by explaining the reason behind the error, without providing the actual solution.

Recent research (Vawter and Martens, 2019) compared 69 different software systems and the type of feedback they provide. The authors found that only 38% of the systems provided explanations when an action in connection with the feedback was required by the user.

A number of studies (Darayani et al., 2018; Ghufron and Rosyida, 2018) examined the effect of feedback generated automatically by different systems, a common target of evaluations being Grammarly[3]. The system currently provides, besides explicit corrections, also feedback consisting of error categories or a short explanation. The results of the investigations analyzing its effectiveness indicate that the automatically generated feedback is useful for students, although the impact of different types of feedback remains unclear.

Besides the type of automatic generated feedback, the type of student revision attempt triggered by these has also been a subject of previous work. These revision attempt types include, for example, additions, replacements, deletions or the absence of a change. Chapelle et al. (2015) found that students disregarded approximately 50% of the feedback generated by the writing evaluation system, Criterion[4], re-

[2] https://www.duolingo.com/
[3] https://www.grammarly.com/
[4] https://criterion.ets.org

gardless of whether it was direct or indirect. This was in part due to the inaccuracy of the generated feedback. Nevertheless, the automatic provision of feedback proved to be useful as it triggered a successful revision in 70% of the cases.

3. Source Corpus

Between 2007 and 2010, City University of Hong Kong hosted a language learning project where English-language tutors reviewed and provided feedback on academic essays written by students, most of whom were native speakers of Chinese. More than 300 TESOL students served as language tutors, and over 4,200 students from a wide range of disciplines took part in the project. Students posted essay drafts as blogs on an e-learning environment called Blackboard Academic Suite. The tutors, which we will refer to as the "teachers" in the rest of this paper, then directly marked linguistic errors or issues on the blogs, using either error categories or open-ended comments. The student essays and teacher feedback have been compiled into a corpus (Lee et al., 2015).

The corpus contains linguistic errors detected and marked by teachers with an error category tag[5] (TAG) or with an open-ended comment (OPEN). An example of each is shown in Table 1, where the location of the error indicated by the teacher is represented by double square brackets in the student response.

Student response	Feedback	Type
The website is, besides [[,]] easy to read.	*Word order*	TAG
[[But]] there still some discrepancies.	*More formal to use "However"*	OPEN

Table 2: Types of teacher feedback in our corpus.

4. Data Extraction

To identify the relevant subset of data to work with in order to answer the research questions posed, a number of filtering and selection steps were implemented. Our aim was to collect instances consisting of the following information: (i) original student sentence with the relevant error type, (ii) revised student sentence and (iii) teacher feedback (open-ended comment or error tag).

The open-ended comments from the source corpus contain some noise due to challenges of data extraction from various file formats and some inconsistency in teachers' practice to indicate the location of errors. Determining revision attempt and outcome automatically based on word-level alignments between original and corrected student sentences would therefore not have been possible with a sufficiently high accuracy. To reduce noise, we filter both teacher comments and students' original and revised sentences for well-formedness as described in the next subsection. For handling the inconsistency of locating errors, revision outcome is manually annotated as detailed in Section 5.

[5]For a complete list of these error categories, see the appendix of Lee et al. (2015).

4.1. Automatic Filtering

First, we perform a generic filtering step on the data, to automatically reduce the noise and identify instances relevant for answering the research questions posed. Initial student versions without teacher comments and final versions without follow-up revisions are excluded. In order to isolate the effect of a single teacher feedback, we only include instances where only one open-ended comment or error tag occurs in the same *error span*, i.e the group of tokens indicated as error location by teachers. We also limit the error span to a maximum of 10 tokens as longer spans are mainly associated with non-corrective (holistic or summative) feedback not requiring a local revision. Moreover, open-ended comments are filtered for the aspects presented in Table 3.

Name	Usage
LEN	max. 300 characters long (incl. spaces)
ALPHA	ratio of alphabetic characters to non-alphabetic ones is min. 0.4
REGEX	does not match certain regular expressions or phrases (e.g. *Figure* / *photo* / *see above* / *see* followed by one or more digits, dates)
NON-CORR	non-corrective feedback starting with *end comment* or *comments:* or being shorter than 50 characters and containing a praise phrase (e.g. *good job, great work*)

Table 3: Filters used for open-ended comments.

These filters aim at excluding cases with data extraction issues (e.g encoding, markup) as well as instances where a local revisions is not expected by the teacher. These are comments where they provide an evaluation of the whole text or when comments contain only reference to figures or previous comments. As establishing revision outcome in these cases would not be relevant, these are not included in our dataset. The student sentences were also checked for the presence of a sufficient amount of alphabetic characters (filter ALPHA from Table 3).

4.2. Error Type Matching

As mentioned above, we focus on errors connected to the use of linking adverbials, words and phrases signaling a logical connection to another sentence(s). The meaning of the connection can be of four different types:

- additive (e.g. *in addition, moreover, as well*),
- adversative (e.g. *however, nonetheless, in contrast*),
- causal (e.g. *therefore, consequently, thus*),
- sequential (e.g. *firstly, finally, to conclude*).

We use the categories and the list of adverbials proposed in Liu (2008, p. 517-518), who includes a comprehensive list (and a straight-forward categorization) of linking adverbials together with frequency information.

We control for error type as confounding factor since some errors may be more difficult or time-consuming for students to correct. Such aspects can influence their willingness to revise errors regardless of the type of teacher feedback. As mentioned in the introduction, linking adverbials are particularly interesting since they are a common error type and they – together with other similar discourse-level errors – are less studied in the SLA literature. Moreover, unlike many other discourse features, identifying open-ended comments targeting the use of linking adverbials is feasible in an automatized way, as described in the remaining part of this section.

Open-ended comments For open-ended comments, we automatically identified linking adverbial errors based on:

(i) explicit terminology used by teachers to describe the phenomenon (e.g. *discourse marker*, *signpost*, *logical link*, *linker*, *joining word*, *conjunct* etc.);

(ii) the occurrence of linking adverbials in quotes in teachers' comments (the requirement of quotes was omitted for adverbials with low frequency in the data);

(iii) the occurrence of linking adverbials in students' sentences when these occurred in tokens highlighted as errors.

Error tags For errors marked with an error tag as feedback, we applied (iii) from above and we restricted the error categories to a subset relevant to linking adverbials, which included the following error categories from the source corpus: *Adverb needed - Part of speech Incorrect*, *Coherence - signposting*, *Coherence - logical sequence*, *Conjunction - Wrong Use*, *Conjunction Missing*, *Conjunction missing OR wrong use*, *Delete this (unnecessary)*, *Word choice*, *Word choice - Level of formality*, *Word order*.

4.3. Manual Filtering

The automatic filtering and error-type matching identified 2,353 cases of interest with student errors and teacher feedback. This was further filtered manually to eliminate all cases that might be false positives for linking adverbial errors due to a misleading use of terminology (e.g. the term *linking word* used to refer to the preposition *by*) or polysemy (e.g. *so* used as intensifier rather than a linking adverbial, e.g. *so well*). Any remaining cases with data extraction issues slipping through the automatic filtering resulting in unreadable or hard-to-interpret teacher comments or student sentences were also eliminated manually. Table 4 shows the size of the resulting dataset of instances consisting of teacher feedback provided for linking adverbial errors.

OPEN	TAG	Total
616	515	**1,131**

Table 4: Number of instances per feedback category.

Furthermore, as a pre-annotation step for revision outcome, we identified the type of revision attempt since a lack of attempt needs no further analysis of success. The most obvious of such cases are when the original and revised student responses are identical (SAME). We focus on the part of the sentence relevant for linking adverbials, other parts of the sentence not relevant to the error may contain changes. Another case is when students remove all or most of an original sentence (REM). Cases where the whole original sentence was missing or where more than two-thirds of the words were revised, were not further assessed for revision outcome. We present the distribution of revision attempt types in Table 5.

Revised	SAME	REM	Total
904	133	94	**1,131**

Table 5: Number of instances per revision attempt type.

As Table 5 shows, the majority of errors (80%) with teacher feedback were addressed by students.

5. Data Annotation

The collected data was manually annotated for feedback directness and revision outcome. The annotations were performed by two annotators with background in linguistics who were financially compensated for their services.

5.1. Feedback Directness

First, the directness of feedback has been annotated. To avoid bias effects, annotators were only presented with the teacher feedback, but they did not have access to the original and revised student sentences corresponding to the teacher feedback. The annotation labels for this aspect are presented in Table 6, where the definitions are based on Sheen and Ellis (2011) for direct and indirect feedback.

Label	Name	Usage
DIR	Direct	Provides the correct form, i.e. the actual word(s) to add, change or delete to correct the error. These are often, but not always, indicated in quotes.
IND	Indirect	Teachers indicate that there was an error, but do NOT include a correction explicitly.

Table 6: Annotation label definitions for directness.

Examples of direct feedback include: *use 'however'* and *only one linker is required - either "and" or "as well as" since they mean the same*. Instances of indirect feedback, on the other hand are: *wrong use of linker* and *-Then- is mainly used in the paragraph not at the beginning*. As these examples show, both direct and indirect feedback can, but does not necessarily contain additional explanation about why the error occurred. For feedback in the form of error tags, information about directness has been automatically added. All error tags were considered indirect, except for *Delete this* which was considered direct, since it provides explicit information about which elements to remove based on the original text highlighting provided by teachers which indicates the relevant segment of the text to delete.

A subset of 200 instances was doubly-annotated for directness to measure inter-annotator (IAA) agreement. Percentage agreement was 88.5% while IAA in terms of Cohen's κ (Cohen, 1960) was 0.77, corresponding to substantial agreement. In the whole dataset, the distribution of instances with direct and indirect feedback was 238 and 893 respectively. The low proportion of direct feedback (21%) is in part due to error tag based feedback being almost always indirect. When considering only open-ended comments, a somewhat higher proportion (30%) constituted direct feedback.

5.2. Revision Outcome

For the annotation of this aspect, the focus was assessing whether students managed to correct successfully errors in their texts based on the feedback given to them by the teachers. Annotators had information about where teachers located the error, but were also cautioned that the actual error might occur before or after the indicated position and they were required to take decisions based on the relevant segment. Misleading cases for error location included sentences where an error was signaled at the end of the sentence while referring to a token other than the last.
Besides the sentence(s) marked by teachers as the location of the error, a larger context of an additional sentence before and after was also available and could be consulted if needed for taking a decision. The annotation manual also included an alphabetically ordered list and a categorized list of linking adverbials from Liu (2008, p. 517-518). These lists could be used, for example, for checking whether a student used an adverbial of the same type as the one suggested by the teacher.
The labels used when annotating revision outcome are presented in Table 7. When assessing the success of a revision, minor issues such as punctuation (e.g. missing comma) were accepted as correct unless teachers explicitly requested their correction.
There were a few cases in which teachers (and the English teaching community in general) show some inconsistency in what is considered stylistically incorrect in academic writing. For these cases, we relied on the opinion expressed by the majority of teacher comments. Based on this, the cases that were considered an unsuccessful revision based on the majority of teacher comments, unless a teacher explicitly suggested these words as the correct alternative in their feedback, were:

- starting a sentence with coordinating conjunctions *and*, *but*, *or*, *so*, which were only acceptable for connecting clauses within a sentence in academic writing;
- using *besides* as it was considered too informal in academic writing by most teachers.

To measure IAA, 500 items were doubly-annotated out of which 6 were excluded from the agreement analysis due to missing values. Results showed a percentage agreement of 95.29% and $\kappa = 0.84$ for the four annotation categories, which indicates a high degree of reliability of the annotations. The distribution of revision outcome types is presented in Table 8.

Label	Name	Usage
GOOD	successful revision	The student successfully corrects the indicated error according to the teacher's suggestion. The word(s) in the revised sentence concerning the error should be semantically and grammatically correct and spelled properly (the rest of the sentence may contain other error types).
ALT	alternative solution	The student successfully corrects the indicated error with a solution different form the one indicated by the teacher. The word(s) in the revised sentence concerning the error should be semantically and grammatically correct and spelled properly (the rest of the sentence may contain other error types).
BAD	unsuccessful revision	The student revises the erroneous part of the sentence, but the indicated error is still not corrected. The sentence may be free from other (e.g. grammatical) errors.
UNC	unclear	It is unclear if the revision is successful (a decision cannot be made based on the available context).

Table 7: Annotation label definitions for revision outcome.

GOOD	ALT	BAD	UNC	Total
771	35	82	16	**904**

Table 8: Distribution of revision outcome types.

The numbers indicate that students seldom adopt a revision strategy different from teachers' suggestions. Moreover, unsuccessful revisions were approximately ten times less frequent than successful ones, which confirms the helpfulness of teacher feedback, but which also creates a significant challenge for creating a dataset with more balanced revision outcome distribution, which could ease machine learning based modeling.
Table 9 show some examples of instances annotated for feedback directness and revision outcome.

6. Initial Statistical Explorations

6.1. Effects of Feedback Type

First, we investigate the effect of feedback type (open-ended comments vs. error tags) on revision attempt and success. Table 10 shows the number of instances per feedback category and revision type, where the 16 instances with unclear revision outcomes labeled as UNC were excluded. Since the count of the ALT category remained low

#	Original	Revised	Feedback	Directness	Outcome
1	*[[But]] there still some discrepancies.*	*[[However]] [[,]] there were some discrepancies.*	*More formal to use "However"*	DIR	GOOD
2	*[[Fifthly]], the higher customer satisfaction , the higher customer loyalty.*	*[[Furthermore]], the higher customer satisfaction, the higher customer loyalty.*	*try a different signpost such as 'next'*	DIR	ALT
3	*For collocation, "phones", "calls" and "switchboard" are related to [[telephone]] that are all regularly co-occurring words in the context.*	*Collocation "phones", "calls" and "switchboard" are related to [[telephones]] which all are regularly co-occurring words in the context.*	*add a connector*	IND	BAD

Table 9: Example instances of open-ended comments with annotations from our dataset.

(35 instances in total), these were merged with GOOD.

	SAME	REM	BAD	GOOD	Total
OPEN	72	43	36	453	**604**
TAG	61	51	46	353	**511**
Total	**133**	**94**	**82**	**806**	**1,115**

Table 10: Distribution of revision types per feedback type.

We can observe that the proportion of open-ended comments with successful revisions is 453 (75%), while for error tags it is 353 (69%), that is 6% lower. This suggests that the added flexibility of open-ended comments which leaves room for additional explanations, facilitates successful revisions. In the case of open-ended comments, however, also not revising the original sentence (SAME) was more common, possibly because they require more effort to read. The lack of revisions in general is considerably less frequent (only 12% of all instances) in our dataset than the 50% reported by Chapelle et al. (2015). This is likely due to the difference in the accuracy of the feedback, which was provided manually by teachers in our case, but which was generated automatically in their study.
It is worth noting that students may also revise for reasons other than teachers' feedback requesting revisions. We have not controlled for this variable, but this type of underling noise in the data is expected to be the same across the two compared feedback types.

6.2. Effects of Feedback Characteristics

To investigate the effect of different characteristics of teacher feedback on students' revision outcome, besides the manually annotated aspect of directness, we extracted automatically two additional variables for open-ended comments, namely whether they contain metalinguistic terms and hedging. These aspects have been pointed out as key and influencing characteristics in previous research for other error types (Sheen and Ellis, 2011; Baker and Bricker, 2010).

Metalinguistic feedback Metalinguistic feedback contains typically grammatical and other specialized terms related to different aspects of linguistics and language (e.g. *vowel*, *plural*, *modifier*) which intend to offer an explanation about the nature of the error. We compile a list of 157 metalinguistic terms based on relevant lexical items from error category names found in the source corpus, from most frequent unigrams in the open-ended comments and by consulting Crystal (2011).

Hedging We detect hedges in the open-ended comments with a lemma-based lexical matching with a set of 55 common hedges consisting of single words from Chapter 5 in Hyland (1998). These include verbs (e.g. *indicate*, *seem*, *might*), nouns (e.g. *claim*), adjectives (e.g. *little*, *possible*) and adverbs (e.g. *quite*, *probably*).
To gain insights into the effect of open-ended comment characteristics on revision outcome, we have performed a Z-test for difference of proportions. The results are shown in Table 11. We found no evidence for a significant difference between the proportion of successful and unsuccessful revisions based on directness and the presence of metalinguistic terms, but hedging, more commonly resulting in correct student rewrites, proved to have a significant effect at α level < 0.05.

Var	BAD %	GOOD %	Diff	Z-stat	p
Direct	5.56	12.8	7.25	-1.276	0.202
Metal.	83.33	77.92	5.41	0.758	0.448
Hedge	11.11	27.37	**16.26**	-2.136	**0.033**

Table 11: Z-test results.

A subsequent chi-square (χ^2) test provided a similar indication as Table 12 shows. Hedging showed a significant association to revision success, just slightly above an α level < 0.05.

Var	χ^2	p
Direct	1.024	0.312
Metal.	0.3	0.584
Hedge	**3.761**	**0.052**

Table 12: Chi-square test results.

7. Challenges and Observations

In this section, we summarize our experience when working with open-ended teacher feedback. As studies applying

computational linguistic methods are relatively few in this area, this type of information may prove useful for others working on similar kind of data. We outline here our observations and some aspects we found challenging when working with this data type, many of which stem from a non-standard (or informal) language use, a certain degree of subjectivity or a lack of consensus. Although in the case of open-ended teacher feedback, writers are highly-skilled professionals, the data is often relatively informal and it does not necessarily follow canonical writing conventions. This poses significant challenges for NLP tools developed using standard written language.

Incomplete sentences Teachers' feedback often contains incomplete sentences, which may be in part because these serve well for the purposes of communicating the intended feedback message effectively and possibly also due to time-pressure. Setting a canonical example of language use appears to be less prioritized, potentially because feedback was directed towards advanced L2 speakers in the source corpus.

Spelling errors Apart from incomplete sentences, open-ended teacher comments also contain spelling errors occasionally, both of which can have a negative impact on the accuracy of NLP processing pipelines typically trained on more formal genres (e.g. newspapers, novels etc.). Word embeddings with sub-word information could efficiently handle the problem of out-of-vocabulary tokens (Bojanowski et al., 2017).

Orthographic conventions As also the examples in Section 5.1. show, teachers may use, for instance, different orthographic conventions for quoting, employing as delimiters single, double quotes or dashes.

Locating errors We have also found some individual variation in pinpointing the exact location of an error, especially for errors spanning multiple tokens and an entire sentence. Relying on token-level alignment information may therefore not be sufficient for automatically locating errors while processing such data.

Terminology There might not always be a consensus in the community about the use of certain terminology. This was the case for our focus error type, to which teachers referred in a variety of ways in their feedback, namely *linking adverbials*, *connectors*, *linkers*, *linking words*, *logical linkers*, and *signposting*. All of these appear indeed in the scientific literature as near-synonyms for this linguistic phenomenon.

Correctness As mentioned in section 5.2., depending on the type of error, there might also be a lack consensus on what is considered correct or acceptable in a certain context and for a specific text type, which makes the annotation of successful corrections more challenging.

Asynchronous dialogue Written feedback is often provided in a more colloquial style which aims at establishing an asynchronous dialogue with the student, including the posing of questions. Therefore, oral feedback categories such as clarification request and elicitation would also be applicable sometimes, e.g.: *this entire description is fairly unclear - do you mean 'so that they are the same'*, *Can you think of a linking phrase to use at the beginning of this sentence to make clear the connection between the previous sentence and this one*.

Uneven data distribution Perhaps the greatest challenge from a computational perspective encountered with the type of data under investigation is the highly uneven distribution of successful and unsuccessful revisions. Since feedback is provided by experts with the specific aim of helping students to correctly revise, the probability of a successful revision is rather high.

8. Conclusion and Future Work

We presented an annotation scheme and a dataset of teacher feedback and revised student errors. We also described the results of our initial quantitative investigations about the effect of feedback type and open-ended feedback characteristics. We found that unsuccessful revisions were rare, which is a challenge for statistical and data-driven analyses when addressing this task. Our results show that the use of hedging in open-ended teacher comments has a positive effect on student revisions, although to a small extent.

The presented resource aims at taking concrete steps towards research and development on feedback in educational applications of NLP, as well as providing a quantitative basis of investigations for computational and theoretical SLA. Given that this is a rather unexplored data type in the NLP community, the insights summarized here about working with this type of data could prove valuable for researchers addressing similar lines of study. Our observations indicate that integrating spell-checkers would be beneficial both for aiding teachers while typing their feedback and for NLP tools processing the data generated by them. Furthermore, for e-learning platforms where the collection of teacher feedback is planned, clear guidelines regarding terminology and locating errors would considerably ease subsequent data processing.

Since our analysis covered only a few salient characteristics of feedback, it is likely that other factors may also have an impact on revision outcome. In the future, we plan to annotate additional data for other error types to enable the application of a wider variety of NLP methods, some of which would require larger amounts of data when investigating feedback effects. The developed annotation schemes for directness and revision outcome can be easily applied to subsets of this (or other) feedback data sources involving other error types. The incorporation of some oral feedback categories would also be worth exploring. Moreover, since the results indicate that lexical hedging has an influence on revision success, examining the presence and influence of strategic hedging on revision outcome would also be a compelling further direction to pursue.

9. Acknowledgements

This work was partially funded by a HKSAR UGC Teaching & Learning Grant (Meeting the Challenge of Teaching and Learning Language in the University: Enhancing Linguistic Competence and Performance in English and Chinese) in the 2016-19 Triennium.

10. Bibliographical References

Baker, W. and Bricker, R. H. (2010). The effects of direct and indirect speech acts on native English and ESL speakers' perception of teacher written feedback. *System*, 38(1):75–84.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chapelle, C. A., Cotos, E., and Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3):385–405.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Crystal, D. (2011). *A Dictionary of Linguistics and Phonetics*. The Language Library. Wiley.

Darayani, N. A., Karyuatry, L. L., and Rizqan, M. D. A. (2018). Grammarly as a tool to improve students' writing quality. *Edulitics (Education, Literature, and Linguistics) Journal*, 3(1):36–42.

Gass, S. M. and Mackey, A. (2013). *The Routledge handbook of second language acquisition*. Routledge.

Ghufron, M. A. and Rosyida, F. (2018). The role of Grammarly in assessing English as a foreign language (EFL) writing. *Lingua Cultura*, 12(4):395–403.

Hyland, K. (1998). *Hedging in scientific research articles*. Pragmatics & Beyond New Series. John Benjamins Publishing.

Larsen-Walker, M. (2017). Can data driven learning address L2 writers' habitual errors with English linking adverbials? *System*, 69:26–37.

Lee, J., Yeung, C. Y., Zeldes, A., Reznicek, M., Lüdeling, A., and Webster, J. (2015). CityU Corpus of Essay Drafts of English Language Learners: a Corpus of Textual Revision in Second Language Writing. *Language Resources and Evaluation*, 49(3):659–683.

Lightbown, P. M. and Spada, N. (1990). Focus-on-form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in second language acquisition*, 12(4):429–448.

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. *International journal of corpus linguistics*, 13(4):491–518.

Lyster, R. and Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in second language acquisition*, 19(1):37–66.

Nagata, R., Vilenius, M., and Whittaker, E. (2014). Correcting Preposition Errors in Learner English Using Error Case Frames and Feedback Messages. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 754–764.

Nagata, R. (2019). Toward a Task of Feedback Comment Generation for Writing Learning. In *Proc. EMNLP*, pages 3197–3206.

Rudzewitz, B., Ziai, R., De Kuthy, K., Möller, V., Nuxoll, F., and Meurers, D. (2018). Generating feedback for English foreign language exercises. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–136.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., and Madnani, N. (2018). Second language acquisition modeling. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65.

Sheen, Y. and Ellis, R. (2011). Corrective feedback in language teaching. In *Handbook of Research in Second Language Teaching and Learning*, pages 593–610. New York: Routledge.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *Tesol Quarterly*, 41(2):255–283.

Vawter, L. and Martens, A. (2019). Categorizing software feedback in current language software. In *IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, pages 258–260.