A Corpus of Annotated Revisions for Studying Argumentative Writing

Fan Zhang **Homa B. Hashemi** **Rebecca Hwa** **Diane Litman**
University of Pittsburgh
Pittsburgh, PA, 15260
{zhangfan,hashemi,hwa,litman}@cs.pitt.edu

Abstract

This paper presents ArgRewrite, a corpus of between-draft revisions of argumentative essays. Drafts are manually aligned at the sentence level, and the writer's purpose for each revision is annotated with categories analogous to those used in argument mining and discourse analysis. The corpus should enable advanced research in writing comparison and revision analysis, as demonstrated via our own studies of student revision behavior and of automatic revision purpose prediction.

1 Introduction

Most writing-related natural language processing (NLP) research focuses on the analysis of single drafts. Examples include document-level quality assessment (Attali and Burstein, 2006; Burstein and Chodorow, 1999), discourse-level analysis and mining (Burstein et al., 2003; Falakmasir et al., 2014; Persing and Ng, 2016), and fine-grained error detection (Leacock et al., 2010; Grammarly, 2016). Less studied is the analysis of changes *between drafts* – a comparison of revisions and the properties of the differences. Research on this topic can support applications involving revision analysis (Zhang and Litman, 2015), paraphrase (Malakasiotis and Androutsopoulos, 2011) and correction detection (Swanson and Yamangil, 2012; Xue and Hwa, 2014).

Although there are some corpora resources for NLP research on writing comparisons, most tend to be between individual sentences/phrases for tasks such as paraphrase comparison (Dolan and Brockett, 2005; Tan and Lee, 2014) or grammar error correction (Dahlmeier et al., 2013; Yannakoudakis et al., 2011). In terms of revision analysis, the most relevant work analyzes Wikipedia revisions (Daxenberger and Gurevych, 2013; Bronner and Monz, 2012); however, the domain of Wikipedia is so specialized that the properties of Wikipedia revisions do not correspond well with other kinds of texts.

This work presents the ArgRewrite corpus[1] to facilitate revision analysis research for argumentative essays. The corpus consists of a collection of three drafts of essays written by university students and employees; the drafts are manually aligned at the sentence level, then the purpose of each revision is manually coded using a revision schema closely related to argument mining/discourse analysis. Within the domain of argumentative essays, the corpus will be useful for supporting research in argumentative revision analysis and the application of argument mining techniques. The corpus may also be useful for research on paraphrase comparisons, grammar error correction, and computational stylistics (Popescu and Dinu, 2008; Flekova et al., 2016). In this paper, we present two example uses of our corpus: 1) rewriting behavior data analysis, and 2) automatic revision purpose classification.

2 Corpus Design Decisions

Consider this scenario: Alice begins her social science argumentative essay with the sentence "Electronic communication allows people to make connections beyond physical limits."

An analytical system might (rightly) identify the sentence as the thesis of her essay, and an evaluative system might give the essay a low score due to this sentence's vagueness and a later lack of evidence (though Alice may not know why she received that score).

Now suppose in a revised draft, Alice expanded

[1]The corpus is based on the ArgRewrite system developed in our prior work (Zhang et al., 2016).

the sentence: "Electronic communication allows people to make connections beyond physical ~~lim-its~~ *location and enriches connections that would have been impossible to make otherwise.*"

An analytical system would still identify the sentence as the thesis, and an evaluative system might raise the overall score a little higher. Alice may become satisfied with the increase and move on. However, there is an opportunity lost – neither the analytical nor the evaluative system addressed the quality of her revision.

A revision analysis system might be helpful for Alice because it would link "limits" to "location and ..." and identify the reason why she made the change – perhaps *adding precision.* If Alice had intended her change as a way to add evidential support for her thesis, she would see that her attempt was not as successful as she hoped.

The above scenario highlights the application of a revision analysis system. This paper is about creating a corpus to enable the development of such systems. Because this is a relatively new problem, there are many possible ways for us to design the corpus. Here we discuss some of our decisions.

First, we need to define the unit of revision. The example above illustrates a phrase-aligned revision. While this offers a fairly precise definition of the scope of a revision, it may be difficult to achieve consistent annotations. For example, the changes may not adhere to any syntactic linguistic unit. For this first corpus, we define our unit of revision to be at the sentence level. In other words, even if a pair of sentences contains multiple edits, the entire sentence pair will be annotated as one sentence revision.

Second, we need to define the quality we want to observe about the revision sentence pair. For this first corpus, we focus on recognizing the purpose of the revision, as in the example above. It is a useful property, and it has previously been studied by others in the literature. People have considered both binary purpose categories such as Content vs. Surface (Faigley and Witte, 1981) or Factual vs. Fluency (Bronner and Monz, 2012) as well as more fine-grained categories (Pfeil et al., 2006; Jones, 2008; Liu and Ram, 2009; Daxenberger and Gurevych, 2012; Zhang and Litman, 2015). Our corpus follows the two-tiered schema used by (Zhang and Litman, 2015) (see Section 3.2).

Third, we not only have to decide on the annotation format, we also need to decide how to obtain

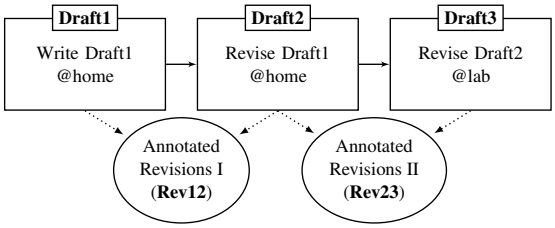


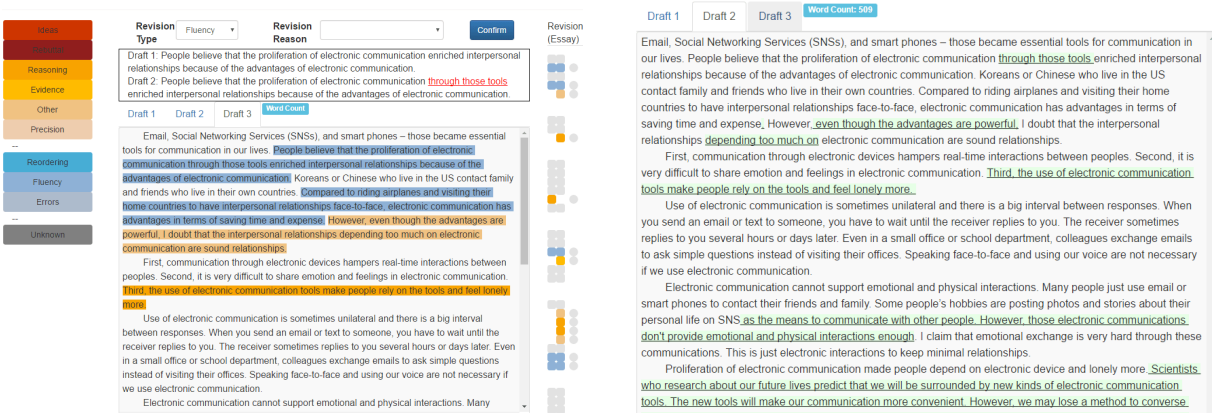
Figure 1: Our collected corpus contains five components: three drafts of an essay and two annotated revisions between drafts.

the raw text: argumentative essays with multiple drafts. We decided to sample from a population of predominantly college students, inclusive of both native and proficient non-native (aka L2) speakers. Comparing to high school students, college students are expected to produce essays having a better organization of the argument elements. Including native and L2 speakers allows for the exploration of possible rewriting differences between writers of varying backgrounds. We decided to give all subjects the same writing prompt and collect three drafts. The identical prompt minimizes the impact of topic difference for argumentation-related study. The collection of three drafts allows for a comparison of revision differences at different stages of rewriting.

Finally, we need a method for eliciting two revised drafts from each writer. Ideally, an instructor would give formative feedback after each draft for each student, but we do not have the resources to carry out such an expensive project. We simulate instructor feedback by asking students to add more examples after the first draft. To elicit a second revised draft, we use two different systems. First, we utilize an idealized[2] version of the ArgRewrite revision analysis system (Zhang et al., 2016). ArgRewrite highlights the locations of revisions at the sentence level and colors the revisions differently according to the revision purpose types. Our second system shows a character-based comparison between subsequent essay drafts[3]. This system is designed to have a similar look as ArgRewrite by highlighting the location of revisions. However, the type of revisions are not provided.

[2]All automatic revision feedback was manually examined/corrected to guarantee correctness.

[3]Code derived from `https://code.google.com/p/google-diff-match-patch/` which implements Myers' algorithm (Myers, 1986).



(a) Interface A. (b) Interface B.

Figure 2: Screenshot of the interfaces. (a) *Interface A* with the annotated revision purposes, (b) *Interface B* with a streamlined character-based diff.

3 The ArgRewrite Corpus

Based on the above design decisions, we have developed a corpus of argumentative essays with three drafts and detailed annotations for sentence-aligned revisions between each consecutive pair of drafts. The main corpus has five elements, with the relationships between them shown in Figure 1; Section 3.1 describes the procedure for obtaining them. Section 3.2 briefly describes the revision schema we used and reports the inter-annotator agreement. Additionally, we have collected meta-data from the participants who contributed to the corpus (discussed in Section 3.3); these data may be useful for user behavior analysis.

3.1 Corpus Development Procedure

We have recruited 60 participants aged 18 years and older, among whom 40 were English native speakers and 20 were non-native speakers with sufficient English proficiency.[4] The study to collect the corpus is carried out in three 40-60 minute sessions over the duration of two weeks.

Draft1 Each participant begins by completing a pre-study questionnaire (Section 3.3) and writing a short essay online. Participants are instructed to keep the essay around 400 words, making a single main point with two supporting examples. They are given the following prompt:

> *"Suppose you've been asked to contribute a short op-ed piece for The New York Times. Argue whether the proliferation of electronic communications (e.g., email, text or other social media) enriches or hinders the development of interpersonal relationships."*

Draft2 A few days later, participants are asked to revise their first draft online based on the following feedback: *Strengthen the essay by adding one more example or reasoning for the claim; then add a rebuttal to an opposing idea; keep the essay at 400 words.* With this feedback we try to push participants to make revisions for later processing by the two interfaces used to create Draft3.

Annotated Revisions I (Rev12) The two drafts are semi-manually aligned at the sentence level.[5] Then, the purpose of each pair of sentence revision is manually coded by a trained annotator, following the annotation guideline (see Section 3.2).

Draft3 Participants write their third draft in a lab environment. This time, they are not given additional instructional feedback. Instead, participants are shown a computer interface that highlights the differences between their first and second drafts. They are asked to revise and create a third draft to improve the general quality of their essay. We experimented with two variations of revision elicitation. Chosen at random, half of the participants (10 L2 participants and 20 Native participants) are shown Interface A, the interface based on the ArgRewrite system (Zhang et al., 2016), which highlights the annotated differences between the drafts (Figure 2(a)); half of the participants are shown In-

[4] i.e., with a TOEFL score higher than 100.

[5] Sentences are first automatically aligned (Zhang and Litman, 2014), then manually corrected by human.

Draft1	Revision Purpose	Draft2	Revision Purpose	Draft3
This world has no restriction on who one can talk to.	Conventions/ Grammar/ Spelling	This world has no restrictions on whom one can talk to.		This world has no restrictions on whom one can talk to.
			Rebuttal/ Reservation	Unfortunately, the younger users of digital communication cannot be entirely protected from the rhetoric of any outsider.
			Warrant/ Reasoning/ Backing	Modern society is now faced with the issue of cyber bullying as a result.
The only aspects of communication that this new development improves are internet navigation and faux internet relatability.	Word-Usage/ Clarity	The only aspects of digital communication that this new development improves are internet navigation and faux internet relatability.	Word-Usage/ Clarity	The only aspects of digital communication that this new development improves are internet navigation and faux internet relationships.
	Claims/ Ideas	Being immersed in the sphere of new technologies can allow for complete isolation from the active, non-digital world.		Being immersed in the sphere of new technologies can allow for complete isolation from the active, non-digital world.

Table 1: Examples from the annotated corpus. The sentences were aligned across the drafts and the revision purposes were labeled on the aligned sentence pairs. From Draft1 to Draft2, there are two *Modify* revisions (*Spelling* and *Clarity*) and one *Add* revision. From Draft2 to Draft3, there are two *Add* revisions (*Rebuttal* and *Reasoning*) and one *Modify* revision (*Clarity*).

terface B, a streamlined character-based diff (Figure 2(b)). In Interface A, some purposes were renamed from the original annotation categories to help the participants better understand the system (as detailed in Table 2)[6]. Both interface groups are asked to read a tutorial about their respective interfaces before beginning to revise. Participants in group A are also asked to verify the manually annotated revision purposes between their first and second drafts. This information is collected to investigate the impact of the difference between the system's recognized and the participant's intended purpose. After completing the final revision, all participants are given a post-study survey about their experiences (Section 3.3). Additionally, participants in group A are asked to verify the automatically predicted revision purposes between their second and third drafts (Section 4.2).

Annotated Revisions II (Rev23) Regardless of which interface the participants used, the second and third draft are compared and annotated by the trained annotator in the same process as before.

[6]Figure 2(a) has two additional categories. Precision was intended to represent revisions that make a sentence more precise. Unknown was intended to represent revisions that cannot be categorized to existing categories. These two categories were not used during annotation as they were reported to be confusing in our pilot studies.

3.2 Revision Annotation Guidelines

Following our prior corpus annotations (Zhang and Litman, 2015), sentence revisions are first coarsely categorized as *Surface* or *Content* changes (Faigley and Witte, 1981), depending on whether any informational content was modified; within each coarse category, we distinguish between several finer categories based on the argumentative and discourse writing literature (Kneupper, 1978; Faigley and Witte, 1981; Burstein et al., 2003). Our adapted schema has three *Surface* categories (Organization, Word Usage/Clarity, and Conventions/Grammar/Spelling) and five *Content* categories (Claim/Ideas, Warrant/Reasoning/Backing, Rebuttal/Reservation, Evidence, and General Content Development). Table 1 shows example aligned sentences in three collected drafts and their annotated revision categories. The edit types of revisions (Add, Delete and Modify) are decided according to the alignment of sentences.

Two annotators (one is experienced, and the other is newly trained) participated in data annotation. The annotators first went through a "training" phase where both annotators annotated 5 files and discussed their disagreements to resolve misunderstandings. Then, both annotators separately annotated 10 new files and Kappa was calculated

Name in Schema	Name in System	Definition
Content	Content	revisions that changed the information of essay
Claims/Ideas	Ideas	revisions that aimed to change the thesis of essay
Warrant/Reasoning/Backing	Reasoning	revisions that aimed to change the reasoning of thesis
Rebuttal/Reservation	Rebuttal	revisions that aimed to change the rebuttal of thesis
Evidence	Evidence	revisions that aimed to change the evidence support for thesis
General Content	Other	other types of content revisions
Surface	Surface	revisions that did not change the information of essay
Organization	Reordering	revisions that switched the order of sentences
Word Usage/Clarity	Fluency	revisions that aimed to make the essay more fluent
Conventions/Grammar/Spelling	Errors	revisions that aimed to fix the spelling/grammar mistakes

Table 2: Definition of category names in Interface A.

L2 (20)	Draft1	Draft2	Draft3
Avg #Words	379.1	412.8	484.7
Avg #Sentences	18.6	20.2	23.7
Avg #Paragraphs	3.9	4.5	4.8
Native (40)	Draft1	Draft2	Draft3
Avg #Words	372.4	394.7	531.6
Avg #Sentences	18.8	20.4	25.8
Avg #Paragraphs	4.0	4.7	5.1

Table 3: Descriptive statistics of the ArgRewrite Corpus, including average number of words, sentences and paragraphs per essay draft.

on the annotation of these 10 new files. The Kappa on this held-out data is 0.84 on the two coarse categories of *Surface* vs. *Content* and 0.71 on the eight fine-grained categories that appear in Table 2. The disagreements between annotators were removed after discussion and the final labels were used as the gold standard annotation.

3.3 Meta-Data

In addition to the raw text and annotations, the corpus release includes participant meta-data from both a pre-study and a post-study survey.

Pre-Study Survey The pre-study survey asks for participants' demographic information as well as their self-reported writing background, such as participants' confidence in their writing ability, the number of drafts they typically make, etc. The questions are listed in Appendix A.

Post-Study Survey The post-study survey contains questions about the participants' in-lab revision experience, such as whether they found the computer interface helpful. All questions are answered on a scale of 1 to 5, ranging from "strongly disagree" to "strongly agree". Details of questions are shown in Appendix B.

3.4 Descriptive Statistics

Table 3 indicates the average number of words/sentences/paragraphs per essay draft. The corpus includes 180 essays: 120 (Draft1 and Draft2) with an average of about 400 words and 60 (Draft3) with an average of around 500 words.

Among the 40 native speakers, there were 29 (72.5%) undergraduates, 6 (15%) graduate students, and 5 (12.5%) non-students (post-docs and lecturers). Among the 20 L2 speakers, there were 4 (20%) undergraduates, and 16 (80%) graduate students; there were 9 Chinese, 2 Bengali, 2 Marathi, 2 Persian, 1 Arabic, 1 Korean, 1 Portuguese, 1 Spanish, and 1 Tamil. In terms of discipline, 33 participants (55%) were from the natural sciences, 24 (40%) from the social sciences, and 2 (3.3%) from the humanities. 1 participant (1.7%) did not reveal his/her discipline.

3.5 Public Release

The corpus is freely available for research usage[7]. The first release includes the raw text plus the revision annotations and the meta-data. The revision annotations are stored as .xlsx files. There are 60 spreadsheet files for revisions from Draft1 to Draft2 and 60 more spreadsheet files for revisions from Draft2 to Draft3. Each spreadsheet file contains two sheets: Old Draft and New Draft. Each row in the sheet represents one sentence in the corresponding draft. The index of the aligned sentence row in the other draft and the type of the revision on the sentence are recorded. The meta-data are in .log text files. Information in the text files are stored using the JSON data format.

4 Example Uses of the Corpus

While the development of a full fledged revision analysis system is outside the scope of this work, we demonstrate potential uses of our corpus with two examples. We first perform statistical analyses on the collected revision data and meta-data

[7] http://argrewrite.cs.pitt.edu

	Content		*Surface*	
	Rev12	Rev23	Rev12	Rev23
L2 (20)	172	78	163	176
Interface A	91	37	71	85
Interface B	81	41	92	91
Native (40)	334	285	303	246
Interface A	177	154	149	111
Interface B	157	131	154	135

Table 4: Number of revisions, by participant groups (language, interface), coarse-grain purposes, and revision drafts (Rev12 is between Draft1-Draft2; Rev23 is between Draft2-Draft3).

to understand aspects of participant behavior. We also use the corpus to train a supervised classifier to automatically predict revision purposes.

4.1 Student Revision Behavior Analysis

While it is well-established that thoughtful revisions improve one's writing, and while many college-level courses require students to submit multiple drafts of writing assignments (Addison and McGee, 2010), instructors rarely monitor and provide feedback to students while they revise. This is partly due to instructors' time constraints and partly due to their uncertainty about how to support students' revisions (Cole, 2014; Melzer, 2014). There is much we do not know about how to stimulate students to self-reflect and revise.

4.1.1 Hypotheses

Using the ArgRewrite Corpus, we can begin to ask and address some questions about revision habits and behaviors. Our first question is: How do different types of revision feedback impact student revision? And relatedly: Does student background (e.g., native vs. L2) make a difference? We thus mine the corpus to test the following hypotheses:

H1. There is a difference in participants' revising behaviors depending on which interface is used to elicit the third draft.

H2. For participants who used Interface A, if the recognized revision purpose differs from the participants' intended revision purpose, participants will further modify their revision.

H3. L2 and native speakers have different behaviors in making revisions.

H1 and **H2** address the first question; **H3** addresses the second.

4.1.2 Methodology

To test the hypotheses, we will use both subjective and objective measures. Subjective measures are based on participant post-study survey answers. Ideally, objective measures should be based on an assessment of improvements in the revised drafts; since we do not have evaluative data at this time, we approximate the degree of improvement using the number of revisions, since these two quantities were demonstrated to be positively correlated (Zhang and Litman, 2015). The objective measures are computed from Tables 4 and 5.

To compare differences between specific subgroups on the subjective and objective measures, we conduct ANOVA tests with two factors. There are multiple factors that can influence the users' rewriting behaviors such as the user's native language, education level and previous revision behaviors, etc. In our study, we try to explore the difference between interface groups considering one of the most salient confounding factors: language. We use one factor as the participant's native language (whether the participant is native or L2) and the other factor as the interface used. To determine correlation between quantitative measures, we conduct Spearman (ρ) and Pearson (r) correlation tests.

4.1.3 Results and Discussion

Testing for H1 Comparing Group A and Group B participants, we observe some differences. First, we detect that Group A agrees with the statement "The system helps me to recognize the weakness of my essay" more so than Group B (Group A has a mean rating of 3.97 ("Agree") while Group B's is 3.17 ("Neutral"), $p < .003$). Second, in Group A, there is a trending positive correlation between the number of revisions[8] from Draft2 to Draft3 and the ratings for the statement "The system encourages me to make more revisions than I usually make" (ρ=.33 and $p < .07$); whereas there is no such correlation for Group B. Additional information about revision purposes may elicit a stronger self-reflection response in Group A participants. In contrast, in Group B, there is a significant negative correlation between the number of Rev12 and ratings for the statement "it is convenient to view my previous revisions with the system" (ρ=-.36 and $p < .05$). This suggests that the character-based interface is ineffective when participants have to reflect on many changes.

[8]The results reported are the normalized numbers $\frac{\#revisions}{\#sentences}$, where #sentences represents the number of sentences in the draft before revision. The absolute numbers were also tested and similar findings were observed.

Revision Purpose	Draft1 to Draft2			Draft2 to Draft3			Totals
	#Add	#Delete	#Modify	#Add	#Delete	#Modify	
Content	294	179	33	320	27	16	869
Claims/Ideas	25	8	4	5	0	0	42
Warrant/Reasoning/Backing	166	83	7	191	13	3	463
Rebuttal/Reservation	23	1	0	13	0	0	37
General Content	50	80	18	86	13	13	260
Evidence	30	7	4	25	1	0	67
Surface	0	0	466	0	0	422	888
Word Usage/Clarity	0	0	362	0	0	357	719
Conventions/Grammar/Spelling	0	0	75	0	0	52	127
Organization	0	0	29	0	0	13	42

Table 5: Number of revisions, by fine-grain revision purposes and edit types (add, delete, modify).

On the other hand, when comparing the number of revisions made by Group A and Group B on Rev23 (controlling for their Rev12 numbers), we did not find a significant difference.

As we did not observe a significant difference in the number of revisions made by the two interface groups, we cannot verify that **H1** is true; possibly a larger pool of participants is needed, or possibly the writing assignment is not extensive enough (in length and in the number of drafts). Another possible explanation is that the system might only motivate the users to make more revisions when the feedback is different from the user's intention. To further verify the correctness of H1, we plan to have the essays graded by experts. The graded scores could allow us to analyze whether essays improved more when Interface A was used.

Testing for H2 Focusing on the 30 participants from Group A, we check the impact of the feedback regarding Rev12 on how they subsequently revise (Rev23). We counted the *Add* and *Modify* revisions where the participant disagrees with the revision purpose assigned by the annotator in Rev12. Of those, we then count the number of times the corresponding sentences were further revised[9]. Of the 53 sentences where the participants disagreed with the annotator, 45 were further revised in the third draft. The ratio is 0.849, much higher than the overall ratio of general Rev12 revisions being further revised in Rev23 (161/394 = 0.409) and the ratio of the agreed Rev12 revisions being revised in Rev23 (67/341 = 0.196). In further analysis, a Pearson correlation test is conducted to check the correlation between the number of Rev23 and the number of disagreements for different types of agreement/disagreements, controlling for the number of Rev12. We find a negative correlation between Rev23 and the number of cases (r=-0.41, $p < .03$) in which the revisions annotated as *Content* are verified by the participants; we also find a positive correlation between Rev23 and the number of cases (r=0.36, $p <= .05$) in which the revisions annotated as *Surface* are intended to be *Content* revisions by the participants. Both findings are consistent with **H2**, suggesting that participants will revise further if they perceive that their intended revisions were not recognized.

Testing for H3 We observe that native and L2 speakers exhibit different behaviors. First, we tested the difference in Content23 and Surface23[10] between these speaker groups with ANOVA. We observe significant difference in the number of content ($p < .02$) and surface ($p < .03$) revisions made by L2 and native speakers. More specifically, our native speakers make more *Content* changes while the L2 speakers make more *Surface* changes. Second, with ANOVA we found a significant interaction effect of the two factors (Group and users' L2 or native status) ($p < .021$) on their ratings for the statement "the system helps me to recognize the weakness of my essay" with L2 speakers having a stronger Interface A preference. Third, we observe a significant positive correlation in the native group between the number of content revisions in Rev23 and the ratings of the statement "the system encourages me to make more revisions than I usually make" (ρ=.4 and $p < .009$). This suggests that giving feedback (from either interface) encourages native speakers to make more content revisions. Finally, in the L2 group, there is a significant negative correlation between the number of surface revisions in Rev12 and the ratings for the statement "the system helps me to recognize the weakness of my es-

[9] *Delete* revisions were ignored as the deleted sentences are not traceable in Draft3

[10] content/surface revisions from Draft2 to Draft3

say" (ρ=-.57 and $p < .008$). This shows that giving feedback to L2 speakers is less helpful when they make more surface revisions. These results are consistent with **H3**.

Summary Our findings suggest that feedback on revisions do impact how students review and rewrite their drafts. However, there are many factors at play, including the interface design and the students' linguistic backgrounds.

4.2 Automatic Revision Identification

Another use of the corpus is to serve as a gold standard for training and testing a revision purpose prediction component for use in an automatic revision analysis system. In the version of ArgRewrite evaluated earlier (Interface A), the manual annotation of revision purposes enabled the system to provide revision feedback to users, which motivated them to improve their writing (**H2**). Automatic argumentative revision purpose prediction has been previously investigated by Zhang and Litman (2015). They developed and reported the performance of a binary classifier for each individual revision category (1 for revisions of the category and 0 for the rest of all revisions) using features from prior research. The availability of our corpus makes it possible for researchers to replicate such methods and conduct further studies.

4.2.1 Hypotheses

In this paper, we repeat the experiment of Zhang and Litman (2015) under different settings to investigate three new hypotheses that can now be investigated given the features of our corpus:

H4. The method used in Zhang and Litman (2015) for high school writings is also useful for the writings of college students.

H5. The same revision classification method works differently for first revision attempts and second revision attempts.

H6. The revision classification model trained on L2 essays has a different preference from the model trained on native essays.

4.2.2 Methodology

We followed the work of (Zhang and Litman, 2015), where unigram features (words) were used as the baseline and the SVM classifier was used. Besides unigrams, three groups of features used in revision analysis, argument mining and discourse analysis research were extracted (*Location*, *Textual* and *Language*) as in Table 6 (Bronner and Monz, 2012; Daxenberger and Gurevych, 2013; Burstein et al., 2001; Falakmasir et al., 2014).

For **H4**, 10-fold (participant) cross-validation is conducted on all the essays in the corpus. Unweighted average F-score for each revision category is reported, using unigram features versus using all features. Zhang and Litman (2015) observed a significant improvement over the unigram baseline using all the features. If **H4** is true, we should expect a similar improvement over the unigram baseline using our corpus.

For **H5**, 10-fold cross-validation was conducted for the revisions from Draft1 to Draft2 and revisions from Draft2 to Draft3 separately. We compared the improvement ratio brought by the advanced features over the unigram baseline.

For **H6**, we trained two classifiers separately with L2 and native essays with all the features. 20 native participants were first randomly selected as the test data. Afterwards classifiers were trained separately using the 20 L2 participants' essays and the remaining 20 native participants' essays. We would expect that the performance of the two trained classifiers is different on the same test data.

4.2.3 Results and Discussion

The first two rows of Table 7 support **H4**. We observe that the method (SVM + all features) used in Zhang and Litman (2015) significantly improves performance (compared to a unigram baseline) for half of the classification tasks, which is similar to Zhang and Litman's results on high school (primarily L1) writing. In our corpus, performance on *Claim*, *Evidence*, *Rebuttal* and *Organization* was not significantly better than the baseline, possibly due to the limited number of positive training samples for these categories (Table 5). For example, one reason that the performance in Table 7 for *Evidence* might be low is that there are less than 100 Evidence instances in Table 5.

For **H5**, the four rows in the middle of Table 7 show the difference of the cross-validation results on first attempt revisions and second attempt revisions. The earlier results using all the revisions, versus now just using only Rev12 or Rev23 revisions are similar, which provides little support for **H5**. With one exception, the features proposed in Zhang and Litman (2015) could again significantly improve the performance over the unigram baseline, for the same set of categories as when using all the revisions. However, for the *Conventions/Grammar/Spelling* category, we did not ob-

Group	Illustration
Location	The location of revised sentences in the paragraph/essay (e.g., whether the sentence is the first or last sentence of the paragraph/essay, the index of the sentence in the paragraph)
Textual	The textual features of revised sentences (e.g., whether the sentence contains a named entity, certain discourse markers ("because", "due to", etc), sentence difference (edit distance, difference in punctuations, etc.) and edit types (Add, Delete or Modify))
Language	The language features of revised sentences (e.g., difference in POS tags, spelling/grammar mistakes)

Table 6: Illustration of features used in the revision classification study.

Experiments	Text-based					Surface		
	Claim	Warrant	General	Evidence	Rebuttal	Org.	Word	Conv
10fold + All Revs + Unigram	0.49	0.58	0.48	0.49	0.49	0.49	0.73	0.49
10fold + All Revs + All features	0.49	0.77∗	0.55∗	0.50	0.49	0.49	0.86∗	0.62∗
10fold + Rev12 + Unigram	0.50	0.58	0.47	0.50	0.50	0.50	0.57	0.62
10fold + Rev12 + All features	0.50	0.77∗	0.56∗	0.50	0.50	0.50	0.72∗	0.72∗
10fold + Rev23 + Unigram	0.50	0.46	0.53	0.49	0.50	0.50	0.58	0.46
10fold + Rev23 + All features	0.50	0.60∗	0.65∗	0.49	0.50	0.50	0.78∗	0.50
20 L2 (train) + 20 Native (test)	0.50	0.72	0.48	0.49	0.50	0.50	0.83	**0.63**
20 Native (train) + 20 Native (test)	0.50	**0.76**	**0.52**	0.49	0.50	0.50	**0.89**	0.54

Table 7: Average unweighted F-score for each binary classification task. The first 6 rows show the average value of 10-fold cross-validation. $*$ indicates significantly better than unigram baseline ($p < .05$). The last 2 rows show the F-value for training on L2/Native data and testing on Native data. **Bold** indicates larger than the number in the other row.

serve a significant improvement for revisions from Draft2 and Draft3. A possible explanation is that there is a bigger difference in the writers' rewriting behavior from Draft2 to Draft3, which increases the difficulty of prediction.

The last two rows of Table 7 support **H6**. Interestingly, we observe a better performance on *Warrant*, *General* and *Word Usage/Clarity* with a classifier trained and tested using native essays. Perhaps essays of native speakers are more similar to each other when revised along these dimensions. For *Conventions/Grammar/Spelling*, in contrast, the classifier trained on L2 data yields a better performance on the same native test set. This may be because the L2 revisions usually include more spelling/grammar corrections.

5 Conclusion and Future Work

We have presented a new corpus for writing comparison research. Currently the corpus focuses on essay revisions made by both native and L2 college students. In addition to three drafts of essays, we have analyzed the drafts to align semantically similar sentences and to assign revision purposes for each revised aligned sentence pair. We have also conducted two studies to demonstrate the use of the corpus for revision behavior analysis and for automatic revision purpose classification.

While in this paper we explored language as one factor influencing rewriting behavior, our corpus also contains information about other potential factors such as gender and education level which we plan to investigate in the future. We also plan to augment the corpus to support additional types of research on revision analysis. Some potential augmentations include more fine-grained revision categories, revision properties such as statement strength (Tan and Lee, 2014) and quality evaluations, and sub-sentential revision scopes.

Acknowledgments

We want to thank Amanda Godley, Geeta Kothari, and the members of the ArgRewrite group (Reed Armstrong, Nicolo Manfredi and Tazin Afrin) for their helpful feedback and the anonymous reviewers for their suggestions. We also want to thank Adam Hobaugh, Dennis Wakefield and Anthony M Taliani for their assistance in the set up of study environments. This material is based upon work supported by the National Science Foundation under Grant No. 1550635. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This research is also funded by the Learning Research and Development Center of the University of Pittsburgh.

References

Joanne Addison and Sharon James McGee. 2010. Writing in high school/writing in college: Research trends and future directions. *College Composition and Communication* pages 147–179.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pages 356–366.

Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*. pages 68–75.

Jill Burstein, Daniel Marcu, Slava Andreyev, and Martin Chodorow. 2001. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*. pages 98–105.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18(1):32–39.

Daniel Cole. 2014. What if the earth is flat? working with, not against, faculty concerns about grammar in student writing. *The WAC Journal* 25:7–35.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. 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*. pages 22–31.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *COLING*. pages 711–726.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 578–589.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication* pages 400–414.

Mohammad Hassan Falakmasir, Kevin D Ashley, Christian D Schunn, and Diane J Litman. 2014. Identifying thesis and conclusion statements in student essays to scaffold peer review. In *International Conference on Intelligent Tutoring Systems*. pages 254–259.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Short Papers)*. pages 313–319.

Grammarly. 2016. Grammarly. http://www.grammarly.com. [Online; accessed 04-10-2017].

John Jones. 2008. Patterns of revision in online writing a study of Wikipedia's featured articles. *Written Communication* 25(2):262–289.

Charles W Kneupper. 1978. Teaching argument: An introduction to the Toulmin model. *College Composition and Communication* 29(3):237–241.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies* 3(1):1–134.

Jun Liu and Sudha Ram. 2009. Who does what: Collaboration patterns in the Wikipedia and their impact on data quality. In *19th Workshop on Information Technologies and Systems*. pages 175–180.

Prodromos Malakasiotis and Ion Androutsopoulos. 2011. A generate and rank approach to sentence paraphrasing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 96–106.

Dan Melzer. 2014. The connected curriculum: Designing a vertical transfer writing curriculum. *The WAC Journal* 25:78–91.

Eugene W Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica* 1(1-4):251–266.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of NAACL-HLT*. pages 1384–1394.

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-Mediated Communication* 12(1):88–113.

Marius Popescu and Liviu P. Dinu. 2008. Rank distance as a stylistic similarity. In *Coling 2008*. pages 91–94.

Ben Swanson and Elif Yamangil. 2012. Correction detection and error type selection as an ESL educational aid. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 357–361.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of ACL (short paper)*.

Huichao Xue and Rebecca Hwa. 2014. Improved correction detection in revised ESL sentences. In *ACL (2)*. pages 599–604.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. 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-Volume 1*. pages 180–189.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B Hashemi. 2016. Argrewrite: A web-based revision assistant for argumentative writings. *NAACL HLT 2016* page 37.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. pages 149–154.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. pages 133–143.

A Questions of the pre-study survey

1. Is English your native language?
2. (only L2 participants) What is your native language?
3. What is your major? Please select the closest discipline to your major.
    - Natural sciences
    - Social sciences
    - Humanities
4. Are you an undergraduate or graduate student?
5. What is your current year of study?
6. When writing a paper for a class, how many drafts of major revisions do you typically make?
7. Overall, how confident are you with your writings? (Not at all confident, Not very confident, Somewhat confident, confident, Extremely confident)
8. (only L2 participants) Please tell us how comfortable you feel about writing in the English language versus writing in your primary language. (Not at all comfortable, Not very comfortable, Somewhat comfortable, comfortable, Extremely comfortable)
9. What are some of your recent classes that have an intensive writing component to them? How did you do in these classes?
10. What aspects of writing do you think you are good at? e.g. vocabulary choice, clear sentences, writing organization.
11. What aspects of writing do you think you can improve?

B Questions of the post-study survey

1. The system allows me to have a better understanding of my previous revision efforts.
2. It is convenient to view my previous revisions with the system.
3. The system helps me to recognize the weakness of my essay.
4. The system encourages me to make more revisions than I usually make.
5. The system encourages me to think more about making more meaningful changes.
6. Overall the system is helpful to my writing.