NAACL'10 Workshop on Innovative Use of NLP for Building Educational Applications

Annotating ESL Errors: Challenges and Rewards

Alla Rozovskaya and Dan Roth
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{rozovska,danr}@illinois.edu

Abstract

In this paper, we present a corrected and error-tagged corpus of essays written by non-native speakers of English. The corpus contains 63000 words and includes data by learners of English of nine first language backgrounds. The annotation was performed at the sentence level and involved correcting all errors in the sentence. Error classification includes mistakes in preposition and article usage, errors in grammar, word order, and word choice. We show an analysis of errors in the annotated corpus by error categories and first language backgrounds, as well as inter-annotator agreement on the task.

We also describe a computer program that was developed to facilitate and standardize the annotation procedure for the task. The program allows for the annotation of various types of mistakes and was used in the annotation of the corpus.

1 Introduction

Work on automated methods for detecting and correcting context dependent mistakes (e.g., (Golding and Roth, 1996; Golding and Roth, 1999; Carlson et al., 2001)) has taken an interesting turn over the last few years, and has focused on correcting mistakes made by non-native speakers of English. Non-native writers make a variety of errors in grammar and word usage. Recently, there has been a lot of effort on building systems for detecting mistakes in article and preposition usage (DeFelice, 2008; Eeg-Olofsson, 2003; Gamon et al., 2008; Han et al., 2006; Tetreault and Chodorow, 2008b). Izumi et al. (2003) consider several error types, including article and preposition mistakes, made by Japanese learners of English, and Nagata et al. (2006) focus on the errors in mass/count noun distinctions with an application to detecting article mistakes also made by Japanese speakers. Article and preposition mistakes have been shown to be very common mistakes for learners of different first language (L1) backgrounds (Dagneaux et al., 1998; Gamon et al., 2008; Izumi et al., 2004; Tetreault and Chodorow, 2008a), but there is no systematic study of a whole range of errors non-native writers produce, nor is it clear what the distribution of different types of mistakes is in learner language.

In this paper, we describe a corpus of sentences written by English as a Second Language (ESL) speakers, annotated for the purposes of developing an automated system for correcting mistakes in text. Although the focus of the annotation were errors in article and preposition usage, all mistakes in the sentence have been corrected. The data for annotation were taken from two sources: The International Corpus of Learner English (ICLE, (Granger et al., 2002a)) and Chinese Learners of English Corpus (CLEC, (Gui and Yang, 2003)). The annotated corpus includes data from speakers of nine first language backgrounds. To our knowledge, this is the first corpus of non-native English text (learner corpus) of fully-corrected sentences from such a diverse group of learners[1]. The size of the annotated corpus is 63000 words, or 2645 sentences. While a corpus

[1] Possibly, except for the Cambridge Learner Corpus http://www.cambridge.org/elt

of this size may not seem significant in many natural language applications, this is in fact a large corpus for this field, especially considering the effort to correct all mistakes, as opposed to focusing on one language phenomenon. This corpus was used in the experiments described in the companion paper (Rozovskaya and Roth, 2010).

The annotation schema that we developed was motivated by our special interest in errors in article and preposition usage, but also includes errors in verbs, morphology, and noun number. The corpus contains 907 article corrections and 1309 preposition corrections, in addition to annotated mistakes of other types.

While the focus of the present paper is on annotating ESL mistakes, we have several goals in mind. First, we present the annotation procedure for the task, including an error classification schema, annotation speed, and inter-annotator agreement. Second, we describe a computer program that we developed to facilitate the annotation of mistakes in text. Third, having such a diverse corpus allows us to analyze the annotated data with respect to the source language of the learner. We show the analysis of the annotated data through an overall breakdown of error types by the writer's first language. We also present a detailed analysis of errors in article and preposition usage. Finally, it should be noted that there are currently very few annotated learner corpora available. Consequently, systems are evaluated on different data sets, which makes performance comparison impossible. The annotation of the data presented here is available[2] and, thus, can be used by researchers who obtain access to these respective corpora[3].

The rest of the paper is organized as follows. First, we describe previous work on the annotation of learner corpora and statistics on ESL mistakes. Section 3 gives a description of the annotation procedure, Section 4 presents the annotation tool that was developed for the purpose of this project and used in the annotation. We then present error statistics based on the annotated corpus across all error types and separately for errors in article and preposition usage. Finally, in Section 6 we describe how we evaluate inter-annotator agreement and show agreement results for the task.

2 Learner Corpora and Error Tagging

In this section, we review research in the annotation and error analysis of learner corpora. For a review of learner corpus research see, for example, (Díaz-Negrillo, 2006; Granger, 2002b; Pravec, 2002). Comparative error analysis is difficult, as there are no standardized error-tagging schemas, but we can get a general idea about the types of errors prevalent with such speakers. Izumi et al. (2004a) describe a speech corpus of Japanese learners of English (NICT JLE). The corpus is corrected and annotated and consists of the transcripts (2 million words) of the audio-recordings of the English oral proficiency interview test. In the NICT corpus, whose error tag set consists of 45 tags, about 26.6% of errors are determiner related, and 10% are preposition related, which makes these two error types the most common in the corpus (Gamon et al., 2008). The Chinese Learners of English corpus (CLEC, (Gui and Yang, 2003)) is a collection of essays written by Chinese learners of beginning, intermediate, and advanced levels. This corpus is also corrected and error-tagged, but the tagging schema does not allow for an easy isolation of article and preposition errors. The International Corpus of Learner English (ICLE, (Granger et al., 2002a)) is a corpus of argumentative essays by advanced English learners. The corpus contains 2 million words of writing by European learners from 14 mother tongue backgrounds. While the entire corpus is not error-tagged, the French subpart of the corpus along with other data by French speakers of a lower level of proficiency has been annotated (Dagneaux et al., 1998). The most common errors for the advanced level of proficiency were found to be lexical errors (words) (15%), register (10%), articles (10%), pronouns (10%), spelling (8%) , verbs (8%).

In a study of 53 post-intermediate ESOL (migrant) learners in New Zealand (Bitchener et al., 2005), the most common errors were found to be prepositions (29%), articles (20%), and verb tense (22%). Dalgish (1985) conducted a study of errors produced by ESL students enrolled at CUNY. It was found that across students of different first

[2] Details about the annotation are accessible from `http://L2R.cs.uiuc.edu/~cogcomp/`

[3] The ICLE and CLEC corpora are commercially available.

languages, the most common error types among 24 different error types were errors in article usage (28%), vocabulary error (20-25%) (word choice and idioms), prepositions (18%), and verb-subject agreement (15%). He also noted that the speakers of languages without article system made considerably more article errors, but the breakdown of other error types across languages was surprisingly similar.

3 Annotation

3.1 Data Selection

Data for annotation were extracted from the ICLE corpus (Granger et al., 2002a) and CLEC (Gui and Yang, 2003). As stated in Section 2, the ICLE contains data by European speakers of advanced level of proficiency, and the CLEC corpus contains essays by Chinese learners of different levels of proficiency. The annotated corpus includes sentences written by speakers of nine languages: Bulgarian, Chinese, Czech, French, German, Italian, Polish, Russian, and Spanish. About half of the sentences for annotation were selected based on their scores with respect to a 4-gram language model built using the English Gigaword corpus (LDC2005T12). This was done in order to exclude sentences that would require heavy editing and sentences with near-native fluency, sentences with scores too high or too low. Such sentences would be less likely to benefit from a system on preposition/article correction. The sentences for annotation were a random sample out of the remaining 80% of the data.

To collect more data for errors in preposition usage, we also manually selected sentences that contained such errors. This might explain why the proportion of preposition errors is so high in our data.

3.2 Annotation Procedure

The annotation was performed by three native speakers of North American English, one undergraduate and two graduate students, specializing in foreign languages and Linguistics, with previous experience in natural language annotation. A sentence was presented to the annotator in the context of the essay from which it was extracted. Essay context can become necessary, especially for the correction of article errors, when an article is acceptable in the context of a sentence, but is incorrect in the context of the essay. The annotators were also encouraged to propose more than one correction, as long as all of their suggestions were consistent with the essay context.

3.3 Annotation Schema

While we were primarily interested in article and preposition errors, the goal of the annotation was to correct all mistakes in the sentence. Thus, our error classification schema[4], though motivated by our interest in errors in article and preposition usage, was also intended to give us a general idea about the types of mistakes ESL students make. A better understanding of the nature of learners' mistakes is important for the development of a robust automated system that detects errors and proposes corrections. Even when the focus of a correction system is on one language phenomenon, we would like to have information about all mistakes in the context: Error information around the target article or preposition could help us understand how noisy data affect the performance.

But more importantly, a learner corpus with error information could demonstrate how mistakes interact in a sentence. A common approach to detecting and correcting context-sensitive mistakes is to deal with each phenomenon independently, but sometimes errors cannot be corrected in isolation. Consider, for example, the following sentences that are a part of the corpus that we annotated.

1. ”I should know all important aspects of English.” → ”I should know all *of the* important aspects of English.”

2. ”But *some of the* people *thought* about him as a parodist of a rhythm-n-blues singer.” → ”But *some people considered* him to be a parodist of a rhythm-n-blues singer.”

3. ”...to be *a* competent avionics *engineer*...” → ...”to become competent avionics *engineers*...”

4. ”...which reflect a traditional female role and a traditional attitude to *a woman*...” → ”...which reflect a traditional female role and a traditional attitude towards *women*...”

5. ”Marx lived in the epoch when there *were* no *entertainments*.” → ”Marx lived in an era when there *was* no *entertainment*.”

In the examples above, errors interact with one another. In example 1, the context requires a definite article, and the definite article, in turn, calls for the

[4] Our error classification was inspired by the classification developed for the annotation of preposition errors (Tetreault and Chodorow, 2008a).

preposition ”of”. In example 2, the definite article after ”some of” is used extraneously, and deleting it also requires deleting preposition ”of”. Another case of interaction is caused by a word choice error: The writer used the verb ”thought” instead of ”considered”; replacing the verb requires also changing the syntactic construction of the verb complement. In examples 3 and 4, the article choice before the words ”engineer” and ”woman” depends on the number value of those nouns. To correctly determine which article should be used, one needs to determine first whether the context requires a singular noun ”engineer” or plural ”engineers”. Finally, in example 5, the form of the predicate in the relative clause depends on the number value of the noun ”entertainment”.

For the reasons mentioned above, the annotation involved correcting all mistakes in a sentence. The errors that we distinguish are *noun number*, *spelling*, *verb form*, and *word form*, in addition to article and preposition errors . All other corrections, the majority of which are lexical errors, were marked as *word replacement*, *word deletion*, and *word insertion*. Table 1 gives a description of each error type.

4 Annotation Tool

In this section, we describe a computer program that was developed to facilitate the annotation process. The main purpose of the program is to allow an annotator to easily mark the type of mistake, when correcting it. In addition, the tool allows us to provide the annotator with sufficient essay context. As described in Section 3, sentences for annotation came from different essays, so each new sentence was usually extracted from a new context. To ensure that the annotators preserved the meaning of the sentence being corrected, we needed to provide them with the essay context. A wider context could affect the annotator’s decision, especially when determining the correct article choice. The tool allowed us to efficiently present to the annotator the essay context for each target sentence.

Fig. 1 shows the program interface. The sentence for annotation appears in the white text box and the annotator can type corrections in the box, as if working in a word processor environment. Above and below the text box we can see the context boxes, where the rest of the essay is shown. Below the lower context box, there is a list of buttons. The pink buttons and the dark green buttons correspond to different error types, the pink buttons are for correcting article and preposition errors, and the dark green buttons – for correcting other errors. The annotator can indicate the type of mistake being corrected by placing the cursor after the word that contains an error and pressing the button that corresponds to this error type. Pressing on an error button inserts a pair of delimiters after the word. The correction can then be entered between the delimiters. The yellow buttons and the three buttons next to the pink ones are the shortcuts that can be used instead of typing in articles and common preposition corrections. The button *None* located next to the article buttons is used for correcting cases of articles and prepositions used superfluously. To correct other errors, the annotator needs to determine the type of error, insert the corresponding delimiters after the word by pressing one of the error buttons and enter the correction between the delimiters.

The annotation rate for the three annotators varied between 30 and 40 sentences per hour.

Table 2 shows sample sentences annotated with the tool. The proposed corrections are located inside the delimiters and follow the word to which the correction refers. When replacing a sequence of words, the sequence was surrounded with curly braces. This is useful if a sequence is a multi-word expression, such as *at last*.

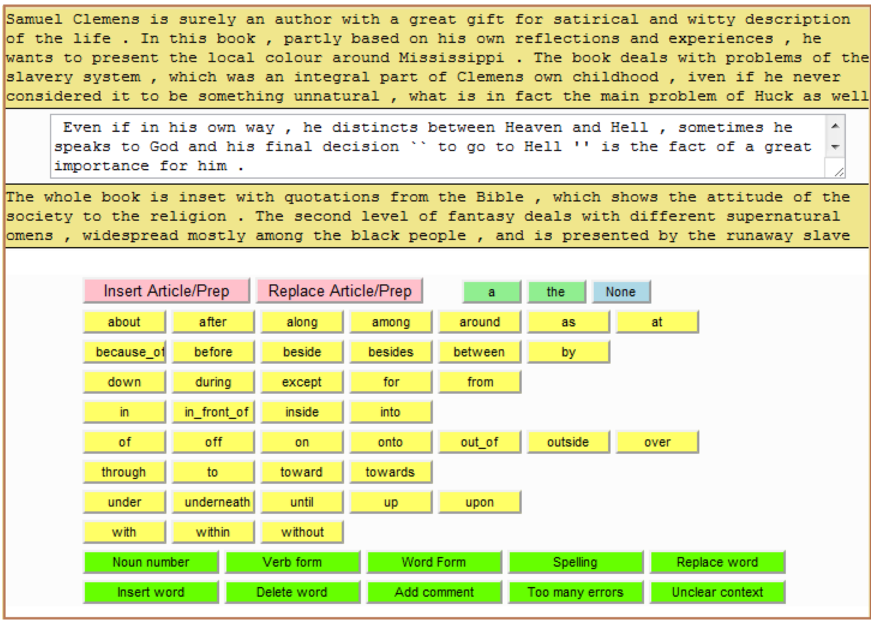
5 Annotation Statistics

In this section, we present the results of the annotation by error type and the source language of the writer.

Table 3 shows statistics for the annotated sentences by language group and error type. Because the sub-corpora differ in size, we show the number of errors per hundred words. In total, the annotated corpus contains 63000 words or 2645 sentences of learner writing. Category *punctuation* was not specified in the annotation, but can be easily identified and includes insertion, deletion, and replacement of punctuation marks. The largest error category is *word replacement*, which combines deleted, inserted words and word substitutions. This is followed by

Error type	Description	Examples
Article error	Any error involving an article	”Women were indignant at [None/the] inequality from men.”
Preposition error	Any error involving a preposition	”...to change their views [to/for] the better.”
Noun number	Errors involving plural/singular confusion of a noun	”Science is surviving by overcoming the mistakes not by uttering the [truths/truth] .”
Verb form	Errors in verb tense and verb inflections	”He [write/writes] poetry.”
Word form	Correct lexeme, but wrong suffix	”It is not [simply/simple] to make professional army .”
Spelling	Error in spelling	”...if a person [commited/committed] a crime...”
Word insertion, deletion, or replacement	Other corrections that do not fall into any of the above categories	”There is a [probability/possibility] that today’s fantasies will not be fantasies tomorrow.”

Table 1: Error classification used in annotation



```
Samuel Clemens is surely an author with a great gift for satirical and witty description
of the life . In this book , partly based on his own reflections and experiences , he
wants to present the local colour around Mississippi . The book deals with problems of the
slavery system , which was an integral part of Clemens own childhood , iven if he never
considered it to be something unnatural , what is in fact the main problem of Huck as well
```

```
 Even if in his own way , he distincts between Heaven and Hell , sometimes he
speaks to God and his final decision `` to go to Hell '' is the fact of a great
importance for him .
```

```
The whole book is inset with quotations from the Bible , which shows the attitude of the
society to the religion . The second level of fantasy deals with different supernatural
omens , widespread mostly among the black people , and is presented by the runaway slave
```

Figure 1: Example of a sentence for annotation as it appears in the annotation tool window. The target sentence is shown in the white box. The surrounding essay context is shown in the brown boxes. The buttons appear below the boxes with text: pink buttons (for marking article and preposition errors), dark green (for marking other errors), light green (article buttons) and yellow (preposition buttons).

Annotated sentence	Corrected errors
1. Television becomes their life , and in many cases it replaces their real life /lives/	noun number (*life* → *lives*)
2. Here I ca n’t $help$ but mention that all these people were either bankers or the Heads of companies or something of that kind @nature, kind@.	word insertion (*help*); word replacement (*kind* → *kind, nature*)
3. We exterminated *have exterminated* different kinds of animals	verb form (*exterminated* → *have exterminated*)
4. ... nearly 30000 species of plants are under the <a> serious threat of disappearance |disappearing|	article replacement (*the* → *a*); word form (*disappearance* → *disappearing*)
5. There is &a& saying that laziness is the engine of the <None> progress	article insertion (*a*); article deletion (*the*)
6. ...experience teaches people to strive to <for> the <None> possible things	preposition replacement (*to* → *for*); article deletion (*the*)

Table 2: Examples of sentences annotated using the annotation tool. Each type of mistake is marked using a different set of delimiters. The corrected words are enclosed in the delimiters and follow the word to which the correction refers. In example 2, the annotator preserved the author’s choice *kind* and added a better choice *nature*.

Source language	**Total sent.**	**Total words**	**Errors per 100 words**	**Corrections by Error Type**								
				Articles	**Prepositions**	**Verb form**	**Word form**	**Noun number**	**Word order**	**Spell.**	**Word repl.**	**Punc.**
Bulgarian	244	6197	11.9	10.3%	12.1%	3.5%	3.1%	3.0%	2.0%	5.0%	46.7%	14.2%
Chinese	468	9327	15.1	12.7%	27.2%	7.9%	3.1%	4.6%	1.4%	5.4%	26.2%	11.3%
Czech	296	6570	12.9	16.3%	10.8%	5.2%	3.4%	2.7%	3.2%	8.3%	32.5%	17.5%
French	238	5656	5.8	6.7%	17.4%	2.1%	4.0%	4.6%	3.1%	9.8%	12.5%	39.8%
German	198	5086	11.4	4.0%	13.0%	4.3%	2.8%	1.9%	2.9%	4.7%	15.4%	51.0%
Italian	243	6843	10.6	5.9%	16.6%	6.4%	1.4%	3.0%	2.4%	4.6%	20.5%	39.3%
Polish	198	4642	10.1	15.1%	16.3%	4.0%	1.3%	1.3%	2.3%	2.1%	12.3%	45.2%
Russian	464	10844	13.0	19.2%	17.8%	3.7%	2.5%	2.5%	2.1%	5.0%	28.3%	18.8%
Spanish	296	7760	15.0	11.5%	14.2%	6.0%	3.8%	2.6%	1.6%	11.9%	37.7%	10.7%
All	**2645**	**62925**	**12.2**	**12.5%**	**17.1%**	**5.2%**	**2.9%**	**3.0%**	**2.2%**	**6.5%**	**28.2%**	**22.5%**

Table 3: Error statistics on the annotated data by source language and error type

the *punctuation* category, which comprises 22% of all corrections. About 12% of all errors involve articles, and prepositions comprise 17% of all errors. We would expect the preposition category to be less significant if we did not specifically look for such errors, when selecting sentences for annotation. Two other common categories are *spelling* and *verb form*. *Verb form* combines errors in verb conjugation and errors in verb tense. It can be observed from the table that there is a significantly smaller proportion of article errors for the speakers of languages that have articles, such as French or German. Lexical errors (word replacement) are more common in language groups that have a higher rate of errors per 100 words. In contrast, the proportion of punctuation mistakes is higher for those learners that make fewer errors overall (cf. French, German, Italian, and Polish). This suggests that punctuation errors are difficult to master, maybe because rules of punctuation are not generally taught in foreign language classes. Besides, there is a high degree of variation in the use of punctuation even among native speakers.

5.1 Statistics on Article Corrections

As stated in Section 2, article errors are one of the most common mistakes made by non-native speakers of English. This is especially true for the speakers of languages that do not have articles, but for advanced French speakers this is also a very common mistake (Dagneaux et al., 1998), suggesting that article usage in English is a very difficult language feature to master.

Han et al. (2006) show that about 13% of noun phrases in TOEFL essays by Chinese, Japanese, and Russian speakers have article mistakes. They also show that learners do not confuse articles randomly and the most common article mistakes are omissions and superfluous article usage. Our findings are summarized in Table 4 and are very similar. We also distinguish between the superfluous use of *a* and *the*, we allows us to observe that most of the cases of extraneously used articles involve article *the* for all language groups. In fact, extraneous *the* is the most common article mistake for the majority of our speakers. Superfluous *the* is usually followed by the omission of *the* and the omission of *a*. Another statistic that our table demonstrates and that was shown previously (e.g. (Dalgish, 1985)) is that learners whose first language does not have articles make more article mistakes: We can see from column 3 of the table that the speakers of German, French and Italian are three to four times less likely to make an article mistake than the speakers of Chinese and all of the Slavic languages. The only exception are Spanish speakers. It is not clear whether the higher error rate is only due to a difference in overall language proficiency (as is apparent from the average number of mistakes by these speakers in Table 3) or to other factors. Finally, the last column in the table indicates that confusing articles with pronouns is a relatively common error and on average accounts for 10% of all article mistakes[5]. Current article correction systems do not address this error type.

[5] An example of such confusion is " To pay for *the* crimes, criminals are put in prison", where *the* is used instead of *their*.

Source language	Errors total	Errors per 100 words	Article mistakes by error type						
			Miss. *the*	Miss. *a*	Extr. *the*	Extr. *a*	Confusion	Mult. labels	Other
Bulgarian	76	1.2	9%	25%	41%	3%	8%	1%	13%
Chinese	179	1.9	20%	12%	48%	4%	7%	2%	7%
Czech	138	2.1	29%	13%	29%	9%	7%	4%	9%
French	22	0.4	9%	14%	36%	14%	0%	23%	5%
German	23	0.5	22%	9%	22%	4%	8%	9%	26%
Italian	43	0.6	16%	40%	26%	2%	9%	0%	7%
Polish	71	1.5	37%	18%	17%	8%	11%	4%	4%
Russian	271	2.5	24%	18%	31%	6%	11%	1%	9%
Spanish	134	1.7	16%	10%	51%	7%	3%	1%	10%
All	**957**	**1.5**	**22%**	**16%**	**36%**	**6%**	**8%**	**3%**	**9%**

Table 4: Distribution of article mistakes by error type and source language of the writer. *Confusion* error type refers to confusing articles *a* and *the*. *Multiple labels* denotes cases where the annotator specified more than one article choice, one of which was used by the learner. *Other* refers to confusing articles with possessive and demonstrative pronouns.

5.2 Statistics on Preposition Corrections

Table 5 shows statistics on errors in preposition usage. Preposition mistakes are classified into three categories: *replacements*, *insertions*, and *deletions*. Unlike with article errors, the most common type of preposition errors is confusing two prepositions. This category accounts for more than half of all errors, and the breakdown is very similar for all language groups. The fourth category in the table, *with original*, refers to the preposition usages that were found acceptable by the annotators, but with a better suggestion provided. We distinguish this case as a separate category because preposition usage is highly variable, unlike, for example, article usage. Tetreault and Chodorow (Tetreault and Chodorow, 2008a) show that agreement between two native speakers on a cloze test targeting prepositions is about 76%, which demonstrates that there are many contexts that license multiple prepositions.

Source language	Errors total	Errors per 100 words	Mistakes by error type			
			Repl.	Ins.	Del.	With orig.
Bulgarian	89	1.4	58%	22%	11%	8%
Chinese	384	4.1	52%	24%	22%	2%
Czech	91	1.4	51%	21%	24%	4%
French	57	1.0	61%	9%	12%	18%
German	75	1.5	61%	8%	16%	15%
Italian	120	1.8	57%	22%	12%	8%
Polish	77	1.7	49%	18%	16%	17%
Russian	251	2.3	53%	21%	17%	9%
Spanish	165	2.1	55%	20%	19%	6%
All	**1309**	**2.1**	**54%**	**21%**	**18%**	**7%**

Table 5: Distribution of preposition mistakes by error type and source language of the writer. *With orig* refers to prepositions judged as acceptable by the annotators, but with a better suggestion provided.

6 Inter-annotator Agreement

Correcting non-native text for a variety of mistakes is challenging and requires a number of decisions on the part of the annotator. Human language allows for many ways to express the same idea. Furthermore, it is possible that the corrected sentence, even when it does not contain clear mistakes, does not sound like a sentence produced by a native speaker. The latter is complicated by the fact that native speakers differ widely with respect to what constitutes acceptable usage (Tetreault and Chodorow, 2008a).

To date, a common approach to annotating non-native text has been to use one rater (Gamon et al., 2008; Han et al., 2006; Izumi et al., 2004; Nagata et al., 2006). The output of human annotation is viewed as the gold standard when evaluating an error detection system. The question of reliability of using one rater has been raised in (Tetreault and Chodorow, 2008a), where an extensive reliability study of human judgments in rating preposition usage is described. In particular, it is shown that inter-annotator agreement on preposition correction is low (kappa value of 0.63) and that native speakers do not always agree on whether a specific preposition constitutes acceptable usage.

We measure agreement by asking an annotator whether a sentence corrected by another person is correct. After all, our goal was to make the sentence sound native-like, without enforcing that errors are corrected in the same way. One hundred sentences annotated by each person were selected and the cor-

Agreement set	Rater	Judged correct	Judged incorrect
Agreement set 1	Rater #2	37	63
	Rater #3	59	41
Agreement set 2	Rater #1	79	21
	Rater #3	73	27
Agreement set 3	Rater #1	83	17
	Rater #2	47	53

Table 6: Annotator agreement at the sentence level. The number next to the agreement set denotes the annotator who corrected the sentences on the first pass. *Judged correct* denotes the proportion of sentences in the agreement set that the second rater did not change. *Judged incorrect* denotes the proportion of sentences, in which the second rater made corrections.

Agreement set	Agreement	kappa
Agreement set 1	56%	0.16
Agreement set 2	78%	0.40
Agreement set 3	60%	0.23

Table 7: Agreement at the sentence level. *Agreement* shows how many sentences in each agreement set were assigned to the same category (”correct”, ”incorrect”) for each of the two raters.

rections were applied. This corrected set was mixed with new sentences and given to the other two annotators. In this manner, each annotator received two hundred sentences corrected by the other two annotators. For each pair of the annotators, we compute agreement based on the 100 sentences on which they did a second pass after the initial corrections by the third rater. To compute agreement at the sentence level, we assign the annotated sentences to one of the two categories: ”correct” and ”incorrect”: A sentence is considered ”correct” if a rater did not make any corrections in it on the second pass [6]. Table 6 shows for each agreement set the number of sentences that were corrected on the second pass. On average, 40.8% of the agreement set sentences belong to the ”incorrect” category, but the proportion of ”incorrect” sentences varies across annotators.

We also compute agreement on the two categories, ”correct” and ”incorrect”. The agreement and the kappa values are shown in Table 7. Agreement on the sentences corrected on the second pass varies between 56% to 78% with kappa values ranging from 0.16 to 0.40. The low numbers reflect the difficulty of the task and the variability of the native speakers’ judgments about acceptable usage. In fact, since the annotation requires looking at several phenomena, we can expect a lower agreement, when compared to agreement rate on one language phenomenon. Suppose rater A disagrees with rater B on a given phenomenon with probability 1/4, then, when there are two phenomena, the probability that he will disagree with at least on of them is $1 - 9/16 = 7/16$. And the probability goes down with the number of phenomena.

7 Conclusion

In this paper, we presented a corpus of essays by students of English of nine first language backgrounds, corrected and annotated for errors. To our knowledge, this is the first fully-corrected corpus that contains such diverse data. We have described an annotation schema, have shown statistics on the error distribution for writers of different first language backgrounds and inter-annotator agreement on the task. We have also described a program that was developed to facilitate the annotation process.

While natural language annotation, especially in the context of error correction, is a challenging and time-consuming task, research in learner corpora and annotation is important for the development of robust systems for correcting and detecting errors.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research is partly supported by a grant from the U.S. Department of Education.

References

J. Bitchener, S. Young and D. Cameron. 2005. The Effect of Different Types of Corrective Feedback on ESL Student Writing. *Journal of Second Language Writing*.

A. J. Carlson and J. Rosen and D. Roth. 2001. Scaling Up Context Sensitive Text Correction. *IAAI*, 45–50.

M. Chodorow, J. Tetreault and N-R. Han. 2007. Detection of Grammatical Errors Involving Prepositions. *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

E. Dagneaux, S. Denness and S. Granger. 1998. Computer-aided Error Analysis. *System*, 26:163–174.

[6] We ignore punctuation corrections.

G. Dalgish. 1985. Computer-assisted ESL Research. *CALICO Journal*, 2(2).

G. Dalgish. 1991. Computer-Assisted Error Analysis and Courseware Design: Applications for ESL in the Swedish Context. *CALICO Journal*, 9.

R. De Felice and S. Pulman. 2008. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of COLING-08*.

A. Díaz-Negrillo and J. Fernández-Domínguez. 2006. Error Tagging Systems for Learner Corpora. *RESLA*, 19:83-102.

J. Eeg-Olofsson and O. Knuttson. 2003. Automatic Grammar Checking for Second Language Learners - the Use of Prepositions. In *Nodalida*.

M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. Dolan, D. Belenko and L. Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. *Proceedings of IJCNLP*.

A. R. Golding and D. Roth. 1996. Applying Winnow to Context-Sensitive Spelling Correction. *ICML*, 182–190.

A. R. Golding and D. Roth. 1999. A Winnow based approach to Context-Sensitive Spelling Correction. *Machine Learning*, 34(1-3):107–130.

S. Granger, E. Dagneaux and F. Meunier. 2002. *International Corpus of Learner English*

S. Granger. 2002. A Bird's-eye View of Learner Corpus Research. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Eds. S. Granger, J. Hung and S. Petch-Tyson, Amsterdam: John Benjamins. 3–33.

S. Gui and H. Yang. 2003. *Zhongguo Xuexizhe Yingyu Yuliaohu. (Chinese Learner English Corpus)*. Shanghai Waiyu Jiaoyu Chubanshe. (In Chinese).

N. Han, M. Chodorow and C. Leacock. 2006. Detecting Errors in English Article Usage by Non-native Speakers. *Journal of Natural Language Engineering*, 12(2):115–129.

E. Izumi, K. Uchimoto, T. Saiga and H. Isahara. 2003. Automatic Error Detection in the Japanese Leaners English Spoken Data. *ACL*.

E. Izumi, K. Uchimoto and H. Isahara. 2004. The Overview of the SST Speech Corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners' Errors. *LREC*.

E. Izumi, K. Uchimoto and H. Isahara. 2004. The NICT JLE Corpus: Exploiting the Language Learner's Speech Database for Research and Education. *International Journal of the Computer, the Internet and Management*, 12(2):119–125.

R. Nagata, A. Kawai, K. Morihiro, and N. Isu. 2006. A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. *ACL/COLING*.

N. Pravec. 2002. Survey of learner corpora. *ICAME Journal*, 26:81–114.

A. Rozovskaya and D. Roth 2010. Training Paradigms for Correcting Errors in Grammar and Usage. In *Proceedings of the NAACL-HLT*, Los-Angeles, CA.

J. Tetreault and M. Chodorow. 2008. Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. *COLING Workshop on Human Judgments in Computational Linguistics*, Manchester, UK.

J. Tetreault and M. Chodorow. 2008. The Ups and Downs of Preposition Error Detection in ESL Writing. *COLING*, Manchester, UK.