Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English

Daniel Dahlmeier[1,2] and **Hwee Tou Ng**[2,3] and **Siew Mei Wu**[4]
[1]SAP Technology and Innovation Platform, SAP Singapore
d.dahlmeier@sap.com
[2]NUS Graduate School for Integrative Sciences and Engineering
[3]Department of Computer Science, National University of Singapore
{danielhe,nght}@comp.nus.edu.sg
[4]Centre for English Language Communication, National University of Singapore
elcwusm@nus.edu.sg

Abstract

We describe the NUS Corpus of Learner English (NUCLE), a large, fully annotated corpus of learner English that is freely available for research purposes. The goal of the corpus is to provide a large data resource for the development and evaluation of grammatical error correction systems. Although NUCLE has been available for almost two years, there has been no reference paper that describes the corpus in detail. In this paper, we address this need. We describe the annotation schema and the data collection and annotation process of NUCLE. Most importantly, we report on an unpublished study of annotator agreement for grammatical error correction. Finally, we present statistics on the distribution of grammatical errors in the NUCLE corpus.

1 Introduction

Grammatical error correction for language learners has recently attracted increasing interest in the natural language processing (NLP) community. Grammatical error correction has the potential to create commercially viable software tools for the large number of students around the world who are studying a foreign language, in particular the large number of students of English as a Foreign Language (EFL).

The success of statistical methods in NLP over the last two decades can largely be attributed to advances in machine learning and the availability of large, annotated corpora that can be used to train and evaluate statistical models for various NLP tasks. The biggest obstacle for grammatical error correction has been that until recently, there was no large, annotated corpus of learner text that could have served as a standard resource for empirical approaches to grammatical error correction (Leacock et al., 2010). The existing annotated learner corpora were all either too small or proprietary and not available to the research community. That is why we decided to create the NUS Corpus of Learner English (NUCLE), a large, annotated corpus of learner texts that is freely available for research purposes. The corpus was built in collaboration with the Centre for English Language Communication (CELC) at NUS. NUCLE consists of about 1,400 student essays from undergraduate university students at NUS with a total of over one million words which are completely annotated with error tags and corrections. All annotations and corrections have been performed by professional English instructors. To the best of our knowledge, NUCLE is the first annotated learner corpus of this size that is freely available for research purposes. However, although the NUCLE corpus has been available for almost two years now, there has been no reference paper that describes the details of the corpus. That makes it harder for other researchers to start working with the NUCLE corpus. In this paper, we address this need by giving a detailed description of the NUCLE corpus, including a description of the annotation schema, the data collection and annotation process, and various statistics on the distribution of grammatical errors in the corpus. Most importantly, we report on an unpublished study of annotator agreement for grammatical error correction that was conducted prior to creating

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 22–31,
Atlanta, Georgia, June 13 2013.

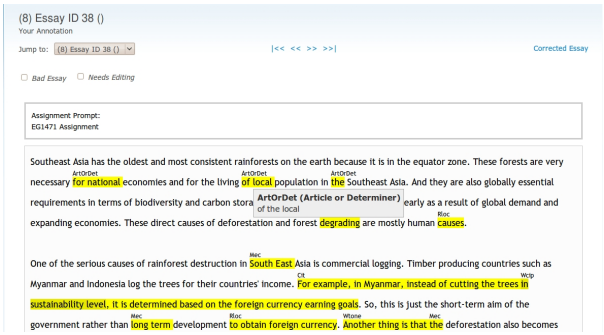

Figure 1: The WAMP annotation interface

the NUCLE corpus. The study gives some insights regarding the difficulty of the annotation task.

The remainder of this paper is organized as follows. The next section explains the annotation schema that was used for labeling grammatical errors. Section 3 reports the results of the inter-annotator agreement study. Section 4 describes the data collection and annotation process. Section 5 contains the error statistics. Section 6 gives the related work, and Section 7 concludes the paper.

2 Annotation Schema

Before starting the corpus creation, we had to develop a set of annotation guidelines. This was done in a pilot study before the actual corpus was created. Three instructors from CELC participated in the pilot study. The instructors annotated a small set of student essays that had been collected by CELC previously. The annotation was performed using an in-house, online annotation tool, called Writing, Annotation, and Marking Platform (WAMP), that was developed by the NUS NLP group specially for creating the NUCLE corpus. The annotation tool allows the annotators to work over the Internet using a web browser. Figure 1 shows a screen shot of the WAMP interface. Annotators can browse through a batch of essays that has been assigned to them and perform the following tasks:

- **Select** arbitrary, contiguous text spans using the cursor to identify grammatical errors.
- **Classify** errors by choosing an error tag from a drop-down menu.
- **Correct** errors by typing the correction into a text box.
- **Comment** to give additional explanations if necessary.

We wanted to impose as few constraints as possible on the annotators and to give them an experience that would closely resemble their usual marking using pen and paper. Therefore, the WAMP annotation tool allows annotators to select arbitrary text spans, including overlapping text spans.

After some annotation trials, we decided to use a tag set which had been developed by CELC in a previous study. Some minor modifications were made to the original tag set based on the feedback of the annotators. The result of the pilot study was a tag set of 27 error categories which are grouped into 13 categories. The tag set is listed in Table 1. It is important to note that our annotation schema not only labels each grammatical error with an error category, but also requires an annotator to provide a suitable correction for the error as well. The annotators were asked to provide a correction that would fix the grammatical error if the selected text span containing the grammatical error is replaced with the correction. If multiple alternative text spans could be selected, the annotators were asked to select the minimal text span so that minimal changes were made to arrive at the corrected text.

We chose to use the tag set in Table 1 since this tag set was developed and used in a previous study at CELC and was found to be a suitable tag set. Furthermore, the tag set offers a reasonable compromise in terms of its complexity. With 27 error categories, it is sufficiently fine-grained to enable meaningful statistics for different error categories, yet not as complex as other tag sets that are much larger in size.

3 Annotator Agreement

How reliably can human annotators agree on whether a word or sentence is grammatically correct? The pilot annotation project gave us the opportunity to investigate this question in a quantitative analysis. Annotator agreement is also a measure for how difficult a task is and serves as a test of whether humans can reliable perform the annotation task with the given tag set. During the pilot study, we randomly sampled 100 essays for measuring annotator agreement. These essays are part of the pilot

Error Tag	Error Category	Description / Example
Verbs		
Vt	Verb Tense	A university **[had conducted — conducted]** the survey last year.
Vm	Verb modal	No one **[will — would]** bother to consider a natural balance.
V0	Missing verb	This **[may — may be]** due to a traditional notion that boys would be the main labor force in a farm family.
Vform	Verb form	Will the child blame the parents after he **[growing — grows]** up?
Subject-verb agreement		
SVA	Subject-verb-agreement	The boy **[play — plays]** soccer.
Articles/determiners		
ArtOrDet	Article or Determiner	From the ethical aspect, sex selection technology should not be used in **[non-medical — a non-medical]** situation.
Nouns		
Nn	Noun Number	Sex selection should therefore be used for medical **[reason — reasons]** and nothing else.
Npos	Noun possessive	The education of **[mother's — mothers]** is a significant factor in reducing son preference.
Pronouns		
Pform	Pronoun form	90% of couples seek treatment for family balancing reasons and 80% of **[those — them]** want girls.
Pref	Pronoun reference	Moreover, children may find it hard to communicate with **[his/her — their]** parents.
Word choice		
Wcip	Wrong collocation/idiom/preposition	Singapore, for example, has invested heavily **[on — in]** the establishment of Biopolis
Wa	Acronyms	Using acronyms without explaining what they stand for.
Wform	Word form	Sex-selection may also result in **[addition — additional]** stress for the family.
Wtone	Tone	**[Isn't it — Is it not]** what you always dreamed for?
Sentence Structure		
Srun	Runons, comma splice	**[Do spare some thought and time, we can make a difference! — Do spare some thought and time. We can make a difference!]** (Should be split into two sentences)
Smod	Dangling modifier	**[Faced — When we are faced]** with the unprecedented energy crisis, finding an alternative energy resource has naturally become the top priority issue.
Spar	Parallelism	The use of sex selection would prevent rather than **[contributing — contribute]** to a distorted sex ratio.
Sfrag	Fragment	Although he is a student from the Arts faculty.
Ssub	Subordinate clause	It is the wrong mindset of people that boys are more superior than girls **[should — that should]** be corrected.

Table 1: NUCLE error categories. Grammatical errors in the examples are printed in bold face in the form **[<mistake>— <correction>]**.

Error Tag	Error Category	Description / Example
Word Order		
WOinc	Incorrect sentence form	Why can **[not we — we not]** choose more intelligent and beautiful babies?
WOadv	Adverb/adjective position	It is similar to the murder of many valuable lives **[only based — based only]** on the couple's own wish.
Transitions		
Trans	Link words/phrases	In the process of selecting the gender of the child, ethical problems arise **[where — because]** many innocent lives of unborn fetuses are taken away.
Mechanics		
Mec	Punctuation, capitalization, spelling, typos	The **[affect — effect]** of that policy has yet to be felt.
Redundancy		
Rloc	Local redundancy	Currently, abortion is available to end a life only **[because of — because]** the fetus or embryo has the wrong sex.
Citation		
Cit	Citation	Poor citation practice.
Others		
Others	Other errors	Any error that does not fit into any other category, but can still be corrected.
Um	Unclear meaning	The quality of the passage is so poor that it cannot be corrected.

Table 1: NUCLE error categories (continued)

data set and are not included in the official NUCLE corpus. The essays were then annotated by our three annotators in a way that each essay was annotated independently by two annotators. Four essays had to be discarded as they were of very poor quality and did not allow for any meaningful correction. This left us with 96 essays with double annotation.

Comparing two sets of annotation is complicated by the fact that the set of annotations that corrects an input text to a corrected output text is ambiguous (Dahlmeier and Ng, 2012). In other words, it is possible that two different sets of annotations produce the same correction. For example, one annotator could choose to select a whole phrase as one error, while the other annotator selects each word individually. Our annotation guidelines ask annotators to select the minimum span that is necessary to correct the error, but we do not enforce any hard constraints and different annotators can have a different perception of where an error starts or ends.

An especially difficult case is the annotation of omission errors, for example missing articles. Selecting a range of whitespace characters is difficult for annotators, especially if the annotation tool is web-based (as whitespace is variable in web pages). We asked annotators to select the previous or next word and include them into the suggested correction. To change *conduct survey* to *conduct **a** survey*, the annotator could change *conduct* to *conduct a*, or change *survey* to *a survey*. If we only compare the exact text spans selected by the annotators when measuring agreement, these different ways to select the context could easily cause us to conclude that the annotators disagree when they in fact agree on the corrected phrase. This would lead to an underestimation of annotator agreement. To address this problem, we perform a simple text span normalization. First, we "grow" the selected context to align with whitespace boundaries. For example, if an annotator just selected the last character *e* of the word *use* and provided *ed* as a correction, we grow this annotation so that the whole word *use* is selected and *used* is the correction. Second, we tokenize the text and "trim" the context by removing tokens at the start and end that are identical in the original and the correction. Finally, the annotations are "projected" onto the individual tokens they span, i.e., an annotation that spans a phrase of multiple to-

Source	:	This phenomenon opposes the real .
Annotator A	:	This phenomenon opposes (the → ϵ (ArtOrDet)) (real → reality (Wform)) .
Annotator B	:	This phenomenon opposes the (real → reality (Wform)) .

Table 2: Example of a sentence from the annotator agreement study with annotations from two different annotators.

kens is broken up into multiple token-level annotations. We align the tokens in the original text span and the tokenized correction string using minimum edit distance. Now, we can compare two annotations in a more meaningful way at the token level. Table 2 shows a tokenized example sentence from the annotator agreement study with annotations from two different annotators. Annotator A and B agree that the first three words *This, phenomenon*, and *opposes* and the final period are correct and do not need any correction. The annotators also agree that the word *real* is part of a word form (Wform) error and should be replaced with *reality*. However, they disagree with respect to the article *the*: annotator A believes there is an article error (ArtOrDet) and that the article has to be deleted while annotator B believes that the article is acceptable in this position.

The example shows that annotator agreement can be measured with respect to three different criteria: whether there is an error, what type of error it is, and how the error should be corrected. Accordingly, we analyze annotator agreement under three different conditions:

- **Identification** Agreement of tagged tokens regardless of error category or correction.
- **Classification** Agreement of error category, given identification.
- **Exact** Agreement of error category and correction, given identification.

In the identification task, we are interested to see how well annotators agree on whether something is a grammatical error or not. In the example above, annotators A and B agree on 5 out of 6 tokens and disagree on one token (*the*). That results in an identification agreement of $5/6 = 83\%$. In the classification task, we investigate how well annotators agree on the type of error, given that both have tagged the token as an error. In the example, the classification agreement is 100% as both annotator A and B tagged the word *real* as a word form (Wform) error. Finally, for the exact task, annotators are considered to agree if they agree on the error category and the correction given that they both have tagged the token as an error. In the example, the exact agreement is 100% as both annotators give the same error category Wform and the same correction *reality* for the word *real*. We use the popular Cohen's Kappa coefficient (Cohen, 1960) to measure annotator agreement between annotators. Cohen's Kappa is defined as

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (1)$$

where $Pr(a)$ is the probability of agreement and $Pr(e)$ is the probability of chance agreement. We can estimate $Pr(a)$ and $Pr(e)$ from the double annotated essays through maximum likelihood estimation. For two annotators A and B, the probability of agreement is

$$Pr(a) = \frac{\text{\#agreed tokens}}{\text{\#total tokens}} \quad (2)$$

where the number of agreed tokens is counted as described above, and the total number of tokens is the total token count of the subset of jointly annotated documents. The probability of chance agreement is computed as

$$\begin{aligned} Pr(e) &= Pr(A=1, B=1) + Pr(A=0, B=0) \\ &= Pr(A=1) \times Pr(B=1) \\ &\quad + Pr(A=0) \times Pr(B=0) \end{aligned}$$

where $Pr(A = 1)$ and $Pr(A = 0)$ symbolize the events of annotator A tagging a token as "error" or "no error" respectively. We make use of the fact that both annotators perform the task independently. $Pr(A = 1)$ and $Pr(A = 0)$ can be computed through maximum likelihood estimation.

$$\begin{aligned} Pr(A=1) &= \frac{\text{\# annotated tokens of annotator A}}{\text{\# total tokens}} \\ Pr(A=0) &= \frac{\text{\# unannotated tokens of annotator A}}{\text{\# total tokens}} \end{aligned}$$

Annotators	Kappa-iden	Kappa-class	Kappa-exact
A – B	0.4775	0.6206	0.5313
A – C	0.3627	0.5352	0.4956
B – C	0.3230	0.4894	0.4246
Average	0.3877	0.5484	0.4838

Table 3: Cohen's Kappa for annotator agreement.

The probabilities $Pr(B = 1)$ and $Pr(B = 0)$ are computed analogously. The chance agreement for this task is quite high, as the number of un-annotated tokens is much higher than the number of annotated tokens. Cohen's Kappa coefficients for the three annotators and the average Kappa coefficient are listed in Table 3. We observe that the Kappa scores are relatively low and that there is a substantial amount of variability in the Kappa coefficients; annotator A and B show a higher agreement with each other than they do with annotator C. According to Landis and Koch (1977), Kappa scores between 0.21 and 0.40 are considered fair, and scores between 0.41 and 0.60 are considered moderate. The average Kappa score for identification can therefore only be considered fair and the Kappa scores for classification and exact agreement are moderate. Thus, an interesting result of the pilot study was that annotators find it harder to agree on whether a word is grammatically correct than agreeing on the type of error or how it should be corrected. The annotator agreement study shows that grammatical error correction, especially grammatical error identification, is a difficult problem.

Our findings support previous research on annotator agreement that has shown that grammatical error correction is a challenging task (Tetreault and Chodorow, 2008; Lee et al., 2009). Tetreault and Chodorow (2008) report a Kappa score of 0.63 which in their words "shows the difficulty of this task and also show how two highly trained raters can produce very different judgments." An interesting related work is (Lee et al., 2009) which investigates the annotation of article and noun number errors. The annotation is performed with either a single sentence context only or the five preceding sentences. The agreement between annotators increases when more context is given, from a Kappa score of 0.55 to a Kappa score of 0.60. Madnani *et al.* (2011) and Tetreault *et al.* (2010) propose crowdsourcing to overcome the problem of annotator variability.

4 Data Collection and Annotation

The main data collection for the NUCLE corpus took place between August and December 2009. We collected a total of 2,249 student essays from 6 English courses at CELC. The courses are designed for students who need language support for their academic studies. The essays were written as course assignments on a wide range of topics, like technology innovation or health care. Some example question prompts are shown in Table 4. All students are at a similar academic level, as they are all undergraduate students at NUS. Students would typically have to write two essay assignments during a course. The length of each essay was supposed to be around 500 words, although most essays were longer than the required length. From this data set, a team of 10 CELC English instructors annotated 1,414 essays with over 1.2 million words between October 2009 and April 2010. Due to budget constraints, we were unfortunately not able to perform double annotations for the main corpus. Annotators were allowed to label an error multiple times if the error could be assigned to more than one error tag, although we observed that annotators did not make much use of this option. Minimal post-processing was done after the annotation process. Annotators were asked to review some corrections that appeared to contain annotation mistakes, for example redundancy errors that did not remove the annotated word. The final results of the annotation exercise were a total of 46,597 error tags. The essays and the annotations were released as the NUCLE corpus through the NUS Enterprise R2M portal in June 2011. The link to the corpus can be found on the NUS NLP group's website[1].

5 NUCLE Corpus Statistics

This section provides basic statistics about the NUCLE corpus and the collected annotations. These statistics already reveal some interesting insights about the nature of grammatical errors in learner text. In particular, we are interested in the following questions: how frequent are errors in the NUCLE corpus and what are the most frequent error

[1] www.comp.nus.edu.sg/∼nlp/corpora.html

"Public spending on the aged should be limited so that money can be diverted to other areas of the country's development." Do you agree?
Surveillance technology such as RFID (radio-frequency identification) should not be used to track people (e.g., human implants and RFID tags on people or products). Do you agree? Support your argument with concrete examples.
Choose a concept or prototype currently in research and development and not widely available in the market. Present an argument on how the design can be improved to enhance safety. Remember to consider influential factors such as cost or performance when you summarize and rebut opposing views. You will need to include very recently published sources in your references.

Table 4: Example question prompts from the NUCLE corpus.

NUS Corpus of Learner English	
Documents	1,414
Sentences	59,871
Word tokens	1,220,257
Word types	30,492
Error annotations	46,597
# of sentences per document	42.34
# of word tokens per document	862.98
# of word tokens per sentence	20.38
# of error annotations per document	32.95
# of error annotations per 100 word tokens	3.82

Table 5: Overview of the NUCLE corpus

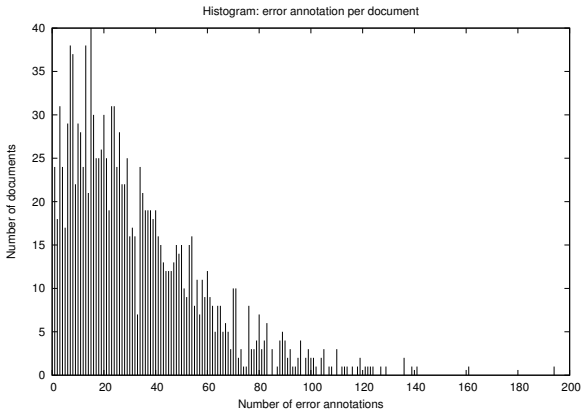

Figure 2: Histogram of error annotations per document in NUCLE.

categories? The basic statistics of the NUCLE corpus are shown in Table 5. In these statistics, we treat multiple alternative annotations for the same error as separate errors, although it could be argued that these should be merged into a single error with multiple alternative corrections. Fortunately, only about 1% of the errors are labeled with more than one annotation. We can see that grammatical errors are very *sparse*, even in learner text. In the NUCLE corpus, there are 46,597 annotated errors for 1,220,257 word tokens. That makes an error density of 3.82 errors per hundred words. In other words, most of the word tokens in the corpus are grammatically correct. This shows that the students whose essays were used for the corpus already have a relative high proficiency of English. When we look at the distribution of errors across documents, we can make another interesting observation. Figure 2 shows a histogram of the number of error annotations per document. The distribution appears non-Gaussian and is heavily skewed to the left with most documents having less than 30 errors while some documents have significantly more errors than the average document. That means that although grammatical errors are rare *in general*, there are also documents with many error annotations. 32 documents have more than 100 error annotations and the highest number of error annotations in a document is 194. The mode, i.e., the most frequent value in the histogram, is 15 which is to the left of the average of 32.95. A similar pattern can be observed when we look at the distribution of errors per sentence. Figure 3 shows a histogram of the number of error annotations per sentence in the NUCLE corpus. For this histogram, only the error annotations which start and end within sentence boundaries are considered (this accounts for 98.6% of all error annotations). Sentence boundaries are determined automatically using the NLTK Punkt sentence splitter[2]. The histogram shows that 57.64% of all sentences have zero errors, 20.48% have exactly one error, and 10.66% have exactly two errors, and 11.21% of all sentences have more than two errors. Although the frequency decreases quickly for higher error counts, the highest observed number of error annotations for a sentence is 28.

[2] `nltk.org`

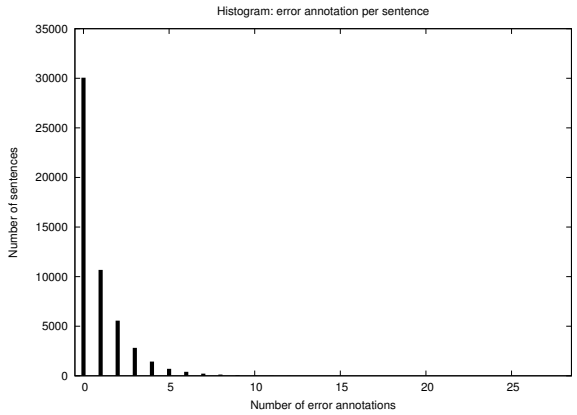

Figure 3: Histogram of error annotations per sentence in NUCLE.

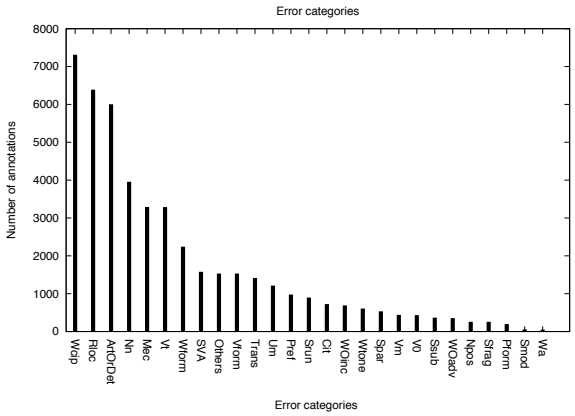

Figure 4: Error categories histogram for the NUCLE corpus.

The skewed distribution of errors in the NUCLE corpus is an interesting observation. A possible explanation for the long tail of the distribution could be a "rich-get-richer" type of dynamics: if a learner has made a lot of mistakes in her essay so far, the chance of her making more errors in the remainder of the essay increases, for example because she makes systematic errors which are likely to be repeated. Explaining the cognitive processes that produce the observed error distribution is beyond the scope of this paper, but it would certainly be an interesting question to investigate.

So far, we have only been concerned with how many errors learners make overall. But it is also important to understand what types of errors language learners make. Error categories that appear more frequently should be addressed with higher priority when creating an automatic error correction system. Figure 4 shows a histogram of error categories. Again, we can observe a skewed distribution with a few error categories being very frequent and many error categories being comparatively infrequent. The top five error categories are wrong collocation/idiom/preposition (Wcip) with 7,312 instances or 15.69% of all annotations, local redundancies (Rloc) (6,390 instances, 13.71%), article or determiner (ArtOrDet) (6,004 instances, 12.88%), noun number (Nn) (3,955 instances, 8.49%), and mechanics (Mec) (3,290 instances, 7.06%). These top five error categories account for 57.83% of all error annotations. The next 5 categories are verb tense (Vt) (3,288 instances, 7.06%) word form (Wform) (2,241 instances, 4.81%), subject-verb agreement (SVA) (1,578 instances, 3.38%), other errors that could not be grouped into any of the error categories (1,532 instances, 3.29%), and Verb form (Vform) (1,531, 3.29%). Together, the top 10 error categories account for 79.66% of all annotated errors. A manual inspection showed that a large percentage of the local redundancy errors involve articles that are deemed redundant by the annotator and should be deleted. These errors could also be considered article or determiner errors. For the Wcip errors, we observed that most Wcip errors are preposition errors. This confirms that articles and prepositions are the two most frequent error categories for EFL learners (Leacock et al., 2010).

6 Related Work

In this section, we compare NUCLE with other learner corpora. While there were almost no annotated learner corpora available for research purposes until recently, non-annotated learner corpora have been available for a while. Two examples are the International Corpus of Learner English (ICLE) (Granger et al., 2002) and the Chinese Learner English Corpus (Gui and Yang., 2003)[3]. Rozovskaya and Roth (2010) annotated a portion of each of these two learner corpora with error categories and corrections. However, with 63,000 words, the annotated data is small compared to NUCLE.

[3]The Chinese Learner English Corpus contains annotations for error types but does not include corrections for the errors.

The Cambridge Learner Corpus (CLC) (Nicholls, 2003) is possibly the largest annotated English learner corpus. Unfortunately, to our knowledge, the corpus is not freely available for research purposes. A subset of the CLC was released in 2011 by Yannakoudakis *et al.* (2011). The released data set contains short essays written by students taking the First Certificate in English (FCE) examination. The data set was also used in the recent HOO 2012 shared task on preposition and determiner correction (Dale et al., 2012). Comparing the essays in the FCE data set and NUCLE, we observe that the essays in the FCE data set are shorter than the essays in NUCLE and show a higher density of grammatical errors. One reason for the higher number of errors (in particular spelling errors) is most likely that the FCE data was not collected from take-home assignments where students have the chance to spell check their writing before submission. But it could also mean that the essays in FCE are from students with a lower proficiency in English compared to NUCLE. With regards to the annotation schema, the CLC annotations include both the type of error (missing, unnecessary, replacement, form) and the part of speech. As a result, the CLC tag set is large with 88 different error categories, far more than the 27 error categories in NUCLE.

Finally, the HOO 2011 shared task (Dale and Kilgarriff, 2011) released an annotated corpus of fragments from academic papers written by non-native speakers and published in a conference or workshop of the Association for Computational Linguistics. The corpus uses the annotation schema from the CLC. Comparing the data set with NUCLE, the HOO 2011 data set is much smaller (about 20,000 words for training and testing, respectively) and represents a specific writing genre (NLP papers). The NUCLE corpus is much larger and covers a broader range of topics.

7 Conclusion

We have presented the NUS Corpus of Learner English (NUCLE), a large, annotated corpus of learner English. The corpus contains over one million words which are completely annotated with grammatical errors and corrections. The NUCLE corpus is freely available for research purposes. We have also reported an inter-annotator agreement study for grammatical error correction. The study shows that grammatical error correction is a difficult task, even for humans. The error statistics from the NUCLE corpus show that learner errors are generally sparse and have a long-tail distribution.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

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

D. Dahlmeier and H.T. Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of HLT-NAACL*, pages 568–572.

R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249.

R. Dale, I. Anisimoff, and G. Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62.

S. Granger, F. Dagneaux, E. Meunier, and M. Paquot. 2002. *The International Corpus of Learner English*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

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

J.R. Landis and G.G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers.

J. Lee, J. Tetreault, and M. Chodorow. 2009. Human evaluation of article and noun number usage: Influences of context and construction variability. In *Proceedings of the Linguistic Annotation Workshop III (LAW3)*, pages 60–63.

N. Madnani, J. Tetreault, M. Chodorow, and R. Rozovskaya. 2011. They can help: using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of ACL:HLT*, pages 508–513.

D. Nicholls. 2003. The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 572–581.

A. Rozovskaya and D. Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

J. Tetreault and M. Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 24–32.

J. Tetreault, E. Filatova, and M. Chodorow. 2010. Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In *Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–48.

H. Yannakoudakis, T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of ACL:HLT*, pages 180–189.