Creating a manually error-tagged and shallow-parsed learner corpus

Ryo Nagata
Konan University
8-9-1 Okamoto,
Kobe 658-0072 Japan
rnagata @ konan-u.ac.jp.

Edward Whittaker **Vera Sheinman**
The Japan Institute for
Educational Measurement Inc.
3-2-4 Kita-Aoyama, Tokyo, 107-0061 Japan
{whittaker,sheinman}@jiem.co.jp

Abstract

The availability of learner corpora, especially those which have been manually error-tagged or shallow-parsed, is still limited. This means that researchers do not have a common development and test set for natural language processing of learner English such as for grammatical error detection. Given this background, we created a novel learner corpus that was manually error-tagged and shallow-parsed. This corpus is available for research and educational purposes on the web. In this paper, we describe it in detail together with its data-collection method and annotation schemes. Another contribution of this paper is that we take the first step toward evaluating the performance of existing POS-tagging/chunking techniques on learner corpora using the created corpus. These contributions will facilitate further research in related areas such as grammatical error detection and automated essay scoring.

1 Introduction

The availability of learner corpora is still somewhat limited despite the obvious usefulness of such data in conducting research on natural language processing of learner English in recent years. In particular, learner corpora tagged with grammatical errors are rare because of the difficulties inherent in learner corpus creation as will be described in Sect. 2. As shown in Table 1, error-tagged learner corpora are very few among existing learner corpora (see Leacock et al. (2010) for a more detailed discussion of learner corpora). Even if data is error-tagged, it is often not available to the public or its access is severely restricted. For example, the Cambridge Learner Corpus, which is one of the largest error-tagged learner corpora, can only be used by authors and writers working for Cambridge University Press and by members of staff at Cambridge ESOL.

Error-tagged learner corpora are crucial for developing and evaluating error detection/correction algorithms such as those described in (Rozovskaya and Roth, 2010b; Chodorow and Leacock, 2000; Chodorow et al., 2007; Felice and Pulman, 2008; Han et al., 2004; Han et al., 2006; Izumi et al., 2003b; Lee and Seneff, 2008; Nagata et al., 2004; Nagata et al., 2005; Nagata et al., 2006; Tetreault et al., 2010b). This is one of the most active research areas in natural language processing of learner English. Because of the restrictions on their availability, researchers have used their own learner corpora to develop and evaluate error detection/correction methods, which are often not commonly available to other researchers. This means that the detection/correction performance of each existing method is not directly comparable as Rozovskaya and Roth (2010a) and Tetreault et al. (2010a) point out. In other words, we are not sure which methods achieve the best performance. Commonly available error-tagged learner corpora are therefore essential to further research in this area.

For similar reasons, to the best of our knowledge, there exists no such learner corpus that is manually shallow-parsed and which is also publicly available, unlike, say, native-speaker corpora such as the Penn Treebank. Such a comparison brings up another crucial question: "Do existing POS taggers and chun-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1210–1219,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

Name	Error-tagged	Parsed	Size (words)	Availability
Cambridge Learner Corpus	Yes	No	30 million	No
CLEC Corpus	Yes	No	1 million	Partially
ETLC Corpus	Partially	No	2 million	Not Known
HKUST Corpus	Yes	No	30 million	No
ICLE Corpus (Granger et al., 2009)	No	No	3.7 million+	Yes
JEFLL Corpus (Tono, 2000)	No	No	1 million	Partially
Longman Learners' Corpus	No	No	10 million	Not Known
NICT JLE Corpus (Izumi et al., 2003a)	Partially	No	2 million	Partially
Polish Learner English Corpus	No	No	0.5 million	No
Janus Pannoius University Learner Corpus	No	No	0.4 million	Not Known

In *Availability*, *Yes* denotes that the full texts of the corpus is available to the public. *Partially* denotes that it is accessible through specially-made interfaces such as a concordancer. The information in this table may not be consistent because many of the URLs of the corpora give only sparse information about them.

Table 1: Learner corpus list.

kers work on learner English as well as on edited text such as newspaper articles?" Nobody really knows the answer to the question. The only exception in the literature is the work by Tetreault et al. (2010b) who evaluated parsing performance in relation to prepositions. Nevertheless, a great number of researchers have used existing POS taggers and chunkers to analyze the writing of learners of English. For instance, error detection methods normally use a POS tagger and/or a chunker in the error detection process. It is therefore possible that a major cause of false positives and negatives in error detection may be attributed to errors in POS-tagging and chunking. In corpus linguistics, researchers (Aarts and Granger, 1998; Granger, 1998; Tono, 2000) use such tools to extract interesting patterns from learner corpora and to reveal learners' tendencies. However, poor performance of the tools may result in misleading conclusions.

Given this background, we describe in this paper a manually error-tagged and shallow-parsed learner corpus that we created. In Sect. 2, we discuss the difficulties inherent in learner corpus creation. Considering the difficulties, in Sect. 3, we describe our method for learner corpus creation, including its data collection method and annotation schemes. In Sect. 4, we describe our learner corpus in detail. The learner corpus is called the Konan-JIEM learner corpus (KJ corpus) and is freely available for research and educational purposes on the web[1]. Another contribution of this paper is that we take the first step toward answering the question about the performance of existing POS-tagging/chunking techniques on learner data. We report and discuss the results in Sect. 5.

2 Difficulties in Learner Corpus Creation

In addition to the common difficulties in creating any corpus, learner corpus creation has its own difficulties. We classify them into the following four categories of the difficulty in:

1. collecting texts written by learners;
2. transforming collected texts into a corpus;
3. copyright transfer; and
4. error and POS/parsing annotation.

The first difficulty concerns the problem in collecting texts written by learners. As in the case of other corpora, it is preferable that the size of a learner corpus be as large as possible where the size can be measured in several ways including the total number of texts, words, sentences, writers, topics, and texts per writer. However, it is much more difficult to create a large learner corpus than to create a

[1] http://www.gsk.or.jp/index_e.html

large native-speaker corpus. In the case of native-speaker corpora, published texts such as newspaper articles or novels can be used as a corpus. By contrast, in the case of learner corpora, we must find learners and then let them write since there are no such published texts written by learners of English (unless they are part of a learner corpus). Here, it should be emphasized that learners often do not spontaneously write but are typically obliged to write, for example, in class, or during an exam. Because of this, learners may soon become tired of writing. This in itself can affect learner corpus creation much more than one would expect especially when creating a longitudinal learner corpus. Thus, it is crucial to keep learners motivated and focused on the writing assignments.

The second difficulty arises when the collected texts are transformed into a learner corpus. This involves several time-consuming and troublesome tasks. The texts must be archived in electronic form, which requires typing every single collected text since learners normally write on paper. Besides, each text must be archived and maintained with accompanying information such as who wrote what text when and on what topic. Optionally, a learner corpus could include other pieces of information such as proficiency, first language, and age. Once the texts have been electronically archived, it is relatively easy to maintain and access them. However, this is not the case when the texts are first collected. Thus, it is better to have an efficient method for managing such information as well as the texts themselves.

The third difficulty concerning copyright is a daunting problem. The copyright for each text must be transferred to the corpus creator so that the learner corpus can be made available to the public. Consider the case when a number of learners participate in a learner corpus creation project and everyone has to sign a copyright transfer form. This issue becomes even more complicated when the writer does not actually have such a right to transfer copyright. For instance, under the Japanese law, those younger than 20 years of age do not have the right; instead their parents do. Thus, corpus creators have to ask learners' parents to sign copyright transfer forms. This is often the case since the writers in learner corpus creation projects are normally junior high school, high school, or college students.

The final difficulty is in error and POS/parsing annotation. For error annotation, several annotation schemes exist (for example, the NICT JLE scheme (Izumi et al., 2005)). While designing an annotation scheme is one issue, annotating errors is yet another. No matter how well an annotation scheme is designed, there will always be exceptions. Every time an exception appears, it becomes necessary to revise the annotation scheme. Another issue we have to remember is that there is a trade-off between the granularity of an annotation scheme and the level of the difficulty in error annotation. The more detailed an annotation scheme is, the more information it can contain and the more difficult identifying errors is, and vice versa.

For POS/parsing annotation, there are also a number of annotation schemes including the Brown tag set, the Claws tag set, and the Penn Treebank tag set. However, none of them are designed to be used for learner corpora. In other words, a variety of linguistic phenomena occur in learner corpora which the existing annotation schemes do not cover. For instance, spelling errors often appear in texts written by learners of English as in *sard year*, which should be *third year*. Grammatical errors prevent us applying existing annotation schemes, too. For instance, there are at least three possibilities for POS-tagging the word *sing* in the sentence *everyone sing together.* using the Penn Treebank tag set: sing/VB, sing/VBP, or sing/VBZ. The following example is more complicated: *I don't success cooking.* Normally, the word *success* is not used as a verb but as a noun. The instance, however, appears in a position where a verb appears. As a result, there are at least two possibilities for tagging: success/NN and success/VB. Errors in mechanics are also problematic as in *Tonight,we* and *beautifulhouse* (missing spaces)[2]. One solution is to split them to obtain the correct strings and then tag them with a normal scheme. However, this would remove the information that spaces were originally missing which we want to preserve. To handle these and other phenomena which are peculiar to learner corpora, we need to develop a novel annotation scheme.

[2]Note that the KJ corpus consists of typed essays.

3 Method

3.1 How to Collect and Maintain Texts Written by Learners

Our text-collection method is based on writing exercises. In the writing exercises, learners write essays on a blog system. This very simple idea of using a blog system naturally solves the problem of archiving texts in electronic form. In addition, the use of a blog system enables us to easily register and maintain accompanying information including who (user ID) writes when (uploaded time) and on what topic (title of blog item). Besides, once registered in the user profile, the optional pieces of information such as proficiency, first language, and age are also easy to maintain and access.

To design the writing exercises, we consulted with several teachers of English and conducted pre-experiments. Ten learners participated in the pre-experiments and were assigned five essay topics on average. Based on the experimental results, we designed the procedure of the writing exercise as shown in Table 2. In the first step, learners are assigned an essay topic. In the second step, they are given time to prepare during which they think about what to write on the given topic before they start writing. We found that this enables the students to write more. In the third step, they actually write an essay on the blog system. After they have finished writing, they submit their essay to the blog system to be registered.

The following steps were considered optional. We implemented an article error detection method (Nagata et al., 2006) in the blog system as a trial attempt to keep the learners motivated since learners are likely to become tired of doing the same exercise repeatedly. To reduce this, the blog system highlights where article errors exist after the essay has been submitted. The hope is that this might prompt the learners to write more accurately and to continue the exercises. In the pre-experiments, the detection did indeed seem to interest the learners and to provide them with additional motivation. Considering these results, we decided to include the fourth and fifth steps in the writing exercises when we created our learner corpus. At the same time, we should of course be aware that the use of error detection affects learners' writing. For example, it may change the distribution of errors. Nagata and Nakatani (2010) reported the effects in detail.

Step	Min.
1. Learner is assigned an essay topic	–
2. Learner prepares for writing	5
3. Learner writes an essay	35
4. System detects errors in the essay	5
5. Learner rewrites the essay	15

Table 2: Procedure of writing exercise.

To solve the problem of copyright transfer, we took legal professional advice but were informed that, in Japan at least, the only way to be sure is to have a copyright transfer form signed every time. We considered having it signed on the blog system, but it soon turned out that this did not work since participating learners may still be too young to have the legal right to sign the transfer. It is left for our long-term future work to devise a better solution to this legal issue.

3.2 Annotation Scheme

This subsection describes the error and POS/chunking annotation schemes. Note that errors and POS/chunking are annotated separately, meaning that there are two files for any given text. Due to space restrictions we limit ourselves to only summarizing our annotation schemes in this section. The full descriptions are available together with the annotated corpus on the web.

3.2.1 Error Annotation

We based our error annotation scheme on that used in the NICT JLE corpus (Izumi et al., 2003a), whose detailed description is readily available, for example, in Izumi et al. (2005). In that annotation scheme and accordingly in ours, errors are tagged using an XML syntax; an error is annotated by tagging a word or phrase that contains it. For instance, a tense error is annotated as follows: *I <v_tns crr="made">make</v_tns> pies last year.*

where *v_tns* denotes a tense error in a verb. It should be emphasized that the error tags contain the information on correction together with error annotation. For instance, *crr="made"* in the above example denotes the correct form of the verb is *made*. For missing word errors, error tags are placed where

a word or phrase is missing (e.g., *My friends live <prp crr="in"></prp> these places.*).

As a pilot study, we applied the NICT JLE annotation scheme to a learner corpus to reveal what modifications we needed to make. The learner corpus consisted of 455 essays (39,716 words) written by junior high and high school students[3]. The following describes the major modifications deemed necessary as a result of the pilot study.

The biggest difference between the NICT JLE corpus and our targeted corpus is that the former is spoken data and the latter is written data. This difference inevitably requires several modifications to the annotation scheme. In speech data, there are no errors in spelling and mechanics such as punctuation and capitalization. However, since such errors are not usually regarded as grammatical errors, we decided simply not to annotate them in our annotation schemes.

Another major difference is fragment errors. Fragments that do not form a complete sentence often appear in the writing of learners (e.g., *I have many books. Because I like reading.*). In written language, fragments can be regarded as a grammatical error. To annotate fragment errors, we added a new tag <f> (e.g., *I have many books.* <f>*Because I like reading.*</f>).

As discussed in Sect. 2, there is a trade-off between the granularity of an annotation scheme and the level of the difficulty in annotating errors. In our annotation scheme, we narrowed down the number of tags to 22 from 46 in the original NICT JLE tag set to facilitate the annotation; the 22 tags are shown in Appendix A. The removed tags are merged into the tag for *other*. For instance, there are only three tags for errors in nouns (number, lexis, and other) in our tag set whereas there are six in the NICT JLE corpus (inflection, number, case, countability, complement, and lexis); the *other* tag (<n_o>) covers the four removed tags.

3.2.2 POS/Chunking Annotation

We selected the Penn Treebank tag set, which is one of the most widely used tag sets, for our POS/chunking annotation scheme. Similar to the error annotation scheme, we conducted a pilot study to determine what modifications we needed to make to the Penn Treebank scheme. In the pilot study, we used the same learner corpus as in the pilot study for the error annotation scheme.

As a result of the pilot study, we found that the Penn Treebank tag set sufficed in most cases except for errors which learners made. Considering this, we determined a basic rule as follows: "Use the Penn Treebank tag set and preserve the original texts as much as possible." To handle such errors, we made several modifications and added two new POS tags (CE and UK) and another two for chunking (XP and PH), which are described below.

A major modification concerns errors in mechanics such as *Tonight,we* and *beautifulhouse* as already explained in Sect. 2. We use the symbol "-" to annotate such cases. For instance, the above two examples are annotated as follows: *Tonight,we/NN-,-PRP* and *beautifulhouse/JJ-NN*. Note that each POS tag is hyphenated. It can also be used for annotating chunks in the same manner. For instance, *Tonight,we* is annotated as *[NP-PH-NP Tonight,we/NN-,-PRP]*. Here, the tag *PH* stands for ϕ chunk label and denotes tokens which are not normally chunked (cf., [NP Tonight/NN] ,/, [NP we/PRP]).

Another major modification was required to handle grammatical errors. Essentially, POS/chunking tags are assigned according to the surface information of the word in question regardless of the existence of any errors. For example, *There is apples.* is annotated as *[NP There/EX] [VP is/VBZ] [NP apples/NNS] ./.* Additionally, we define the CE[4] tag to annotate errors in which learners use a word with a POS which is not allowed such as in *I don't success cooking.* The CE tag encodes a POS which is obtained from the surface information together with the POS which would have been assigned to the word if it were not for the error. For instance, the above example is tagged as *I don't success/CE:NN:VB cooking.* In this format, the second and third POSs are separated by ":" which denotes the POS which is obtained from the surface information and the POS which would be assigned

[3]The learner corpus had been created before this reported work started. Learners wrote their essays on paper. Unfortunately, this learner corpus cannot be made available to the public since the copyrights were not transferred to us.

[4]CE stands for cognitive error.

to the word without an error. The user can select either POS depending on his or her purposes. Note that the CE tag is compatible with the basic annotation scheme because we can retrieve the basic annotation by extracting only the second element (i.e., success/NN). If the tag is unknown because of grammatical errors or other phenomena, UK and XP[5] are used for POS and chunking, respectively.

For spelling errors, the corresponding POS and chunking tag are assigned to mistakenly spelled words if the correct forms can be guessed (e.g., [NP sird/JJ year/NN]); otherwise UK and XP are used.

4 The Corpus

We carried out a learner corpus creation project using the described method. Twenty six Japanese college students participated in the project. At the beginning, we had the students or their parents sign a conventional paper-based copyright transfer form. After that, they did the writing exercise described in Sect. 3 once or twice a week over three months. During that time, they were assigned ten topics, which were determined based on a writing textbook (Okihara, 1985). As described in Sect. 3, they used a blog system to write, submit, and rewrite their essays. Through out the exercises, they did not have access to the others' essays and their own previous essays.

As a result, 233 essays were collected; Table 3 shows the statistics on the collected essays. It turned out that the learners had no difficulties in using the blog system and seemed to focus on writing. Out of the 26 participants, 22 completed the 10 assignments while one student quit before the exercises started.

We annotated the grammatical errors of all 233 essays. Two persons were involved in the annotation. After the annotation, another person checked the annotation results; differences in error annotation were resolved by consulting the first two. The error annotation scheme was found to work well on them. The error-annotated essays can be used for evaluating error detection/correction methods.

Number of essays	233
Number of writers	25
Number of sentences	3,199
Number of words	25,537

Table 3: Statistics on the learner corpus.

[5]UK and XP stand for unknown and X phrase, respectively.

For POS/chunking annotation, we chose 170 essays out of 233. We annotated them using our POS/chunking scheme; hereafter, the 170 essays will be referred to as the shallow-parsed corpus.

5 Using the Corpus and Discussion

5.1 POS Tagging

The 170 essays in the shallow-parsed corpus was used for evaluating existing POS-tagging techniques on texts written by learners. It consisted of 2,411 sentences and 22,452 tokens.

HMM-based and CRF-based POS taggers were tested on the shallow-parsed corpus. The former was implemented using tri-grams by the author. It was trained on a corpus consisting of English learning materials (213,017 tokens). The latter was CRFTagger[6], which was trained on the WSJ corpus. Both use the Penn Treebank POS tag set.

The performance was evaluated using accuracy defined by

$$\frac{\text{number of tokens correctly POS-tagged}}{\text{number of tokens}}. \quad (1)$$

If the number of tokens in a sentence was different in the human annotation and the system output, the sentence was excluded from the calculation. This discrepancy sometimes occurred because the tokenization of the system sometimes differed from that of the human annotators. As a result, 19 and 126 sentences (215 and 1,352 tokens) were excluded from the evaluation in the HMM-based and CRF-based POS taggers, respectively.

Table 4 shows the results. The second column corresponds to accuracies on a native-speaker corpus (sect. 00 of the WSJ corpus). The third column corresponds to accuracies on the learner corpus.

As shown in Table 4, the CRF-based POS tagger suffers a decrease in accuracy as expected. Interestingly, the HMM-based POS tagger performed better on the learner corpus. This is perhaps because it

[6]"CRFTagger: CRF English POS Tagger," Xuan-Hieu Phan, `http://crftagger.sourceforge.net/`, 2006.

was trained on a corpus consisting of English learning materials whose distribution of vocabulary was expected to be relatively similar to that of the learner corpus. By contrast, it did not perform well on the native-speaker corpus because the size of the training corpus was relatively small and the distribution of vocabulary was not similar, and thus unknown words often appeared. This implies that selecting appropriate texts as a training corpus may improve the performance.

Table 5 shows the top five POSs mistakenly tagged as other POSs. An obvious cause of mistakes in both taggers is that they inevitably make errors in the POSs that are not defined in the Penn Treebank tag set, that is, UK and CE. A closer look at the tagging results revealed that phenomena which were common to the writing of learners were major causes of other mistakes. Errors in capitalization partly explain why the taggers made so many mistakes in NN (singular nouns). They often identified erroneously capitalized common nouns as proper nouns as in *This Summer/NNP Vacation/NNP*. Spelling errors affected the taggers in the same way. Grammatical errors also caused confusion between POSs. For instance, omission of a certain word often caused confusion between a verb and an adjective as in *I frightened/VBD.* which should be *I (was) frightened/JJ.* Another interesting case is expressions that learners overuse (e.g., *and/CC so/RB on/RB* and *so/JJ so/JJ*). Such phrases are not erroneous but are relatively infrequent in native-speaker corpora. Therefore, the taggers tended to identify their POSs according to the surface information on the tokens themselves when such phrases appeared in the learner corpus (e.g., *and/CC so/RB* <u>*on/IN*</u> and <u>*so/RB so/RB*</u>). We should be aware that tokenization is also problematic although failures in tokenization were excluded from the accuracies.

The influence of the decrease in accuracy on other NLP tasks is expected to be task and/or method dependent. Methods that directly use or handle sequences of POSs are likely to suffer from it. An example is the error detection method (Chodorow and Leacock, 2000), which identifies unnatural sequences of POSs as grammatical errors in the writing of learners. As just discussed above, existing techniques often fail in sequences of POSs that have a grammatical error. For instance, an existing POS tagger likely tags the sentence *I frightened.* as *I/PRP frightened/VBD ./.* as we have just seen, and in turn the error detection method cannot identify it as an error because the sequence *PRP VBD* is not unnatural; it would correctly detect it if the sentence were correctly tagged as *I/PRP frightened/JJ ./.* For the same reason, the decrease in accuracy may affect the methods (Aarts and Granger, 1998; Granger, 1998; Tono, 2000) for extracting interesting sequences of POSs from learner corpora; for example, BOS[7] PRP JJ is an interesting sequence but is never extracted unless the phrase is correctly POS-tagged. It requires further investigation to reveal how much impact the decrease has on these methods. By contrast, error detection/correction methods based on the bag-of-word features (or feature vectors) are expected to suffer less from it since mistakenly POS-tagged tokens are only one of the features. At the same time, we should notice that if the target errors are in the tokens that are mistakenly POS-tagged, the detection will likely fail (e.g., verbs should be correctly identified in tense error detection).

In addition to the above evaluation, we attempted to improve the POS taggers using the transformation-based POS-tagging technique (Brill, 1994). In the technique, transformation rules are obtained by comparing the output of a POS tagger and the human annotation so that the differences between the two are reduced. We used the shallow-

Method	Native Corpus	Learner Corpus
CRF	0.970	0.932
HMM	0.887	0.926

Table 4: POS-tagging accuracy.

HMM		CRF	
POS	Freq.	POS	Freq.
NN	259	NN	215
VBP	247	RB	166
RB	163	CE	144
CE	150	JJ	140
JJ	108	FW	86

Table 5: Top five POSs mistakenly tagged.

[7] BOS denotes a beginning of a sentence.

Method	Original	Improved
CRF	0.932	0.934
HMM	0.926	0.933

Table 6: Improvement obtained by transformation.

parsed corpus as a test corpus and the other manually POS-tagged corpus created in the pilot study described in Subsect. 3.2.1 as a training corpus. We used POS-based and word-based transformations as Brill (1994) described.

Table 6 shows the improvements together with the original accuracies. Table 6 reveals that even the simple application of Brill's technique achieves a slight improvement in both taggers. Designing the templates of the transformation for learner corpora may achieve further improvement.

5.2 Head Noun Identification

In the evaluation of chunking, we focus on head noun identification. Head noun identification often plays an important role in error detection/correction. For example, it is crucial to identify head nouns to detect errors in article and number.

We again used the shallow-parsed corpus as a test corpus. The essays contained 3,589 head nouns. We implemented an HMM-based chunker using 5-grams whose input is a sequence of POSs, which was obtained by the HMM-based POS tagger described in the previous subsection. The chunker was trained on the same corpus as the HMM-based POS tagger. The performance was evaluated by recall and precision defined by

$$\frac{\text{number of head nouns correctly identified}}{\text{number of head nouns}} \quad (2)$$

and

$$\frac{\text{number of head nouns correctly identified}}{\text{number of tokens identified as head noun}}, \quad (3)$$

respectively.

Table 7 shows the results. To our surprise, the chunker performed better than we had expected. A possible reason for this is that sentences written by learners of English tend to be shorter and simpler in terms of their structure.

The results in Table 7 also enable us to quantitatively estimate expected improvement in error detection/correction which is achieved by improving chunking. To see this, let us define the following symbols: r: Recall of head noun identification, R: recall of error detection without chunking error, $\hat{R}$ recall of error detection with chunking error. R and $\hat{R}$ are interpreted as the true recall of error detection and its observed value when chunking error exists, respectively. Here, note that $\hat{R}$ can be expressed as $\hat{R} = rR$. For instance, according to Han et al. (2006), their method achieves a recall of 0.40 (i.e., $\hat{R} = 0.4$), and thus $R = 0.44$ assuming that chunking errors exist and recall of head noun identification is $r = 0.90$ just as in this evaluation. Improving r to $r = 0.97$ would achieve $R = 0.43$ without any modification to the error detection method. Precision can also be estimated in a similar manner although it requires a more complicated calculation.

6 Conclusions

In this paper, we discussed the difficulties inherent in learner corpus creation and a method for efficiently creating a learner corpus. We described the manually error-annotated and shallow-parsed learner corpus which was created using this method. We also showed its usefulness in developing and evaluating POS taggers and chunkers. We believe that publishing this corpus will give researchers a common development and test set for developing related NLP techniques including error detection/correction and POS-tagging/chunking, which will facilitate further research in these areas.

A Error tag set

This is the list of our error tag set. It is based on the NICT JLE tag set (Izumi et al., 2005).

- n: noun
  - num: number
  - lxc: lexis
  - o: other
- v: verb
  - agr: agreement

Recall	Precision
0.903	0.907

Table 7: Performance on head noun identification.

  - tns: tense
  - lxc: lexis
  - o: other
- mo: auxiliary verb
- aj: adjective
  - lxc: lexis
  - o: other
- av: adverb
- prp: preposition
  - lxc: lexis
  - o: other
- at: article
- pn: pronoun
- con: conjunction
- rel: relative clause
- itr: interrogative
- olxc: errors in lexis in more than two words
- ord: word order
- uk: unknown error
- f: fragment error

References

Jan Aarts and Sylviane Granger. 1998. *Tag sequences in learner corpora: a key to interlanguage grammar and discourse*. Longman Pub Group, London.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proc. of 12th National Conference on Artificial Intelligence*, pages 722–727.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proc. of 1st Meeting of the North America Chapter of ACL*, pages 140–147.

Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proc. of 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30.

Rachele De Felice and Stephen G. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proc. of 22nd International Conference on Computational Linguistics*, pages 169–176.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2*. Presses universitaires de Louvain.

Sylviane Granger. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. P. Cowie, editor, *Phraseology: theory, analysis, and application*, pages 145–160. Clarendon Press.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2004. Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proc. of 4th International Conference on Language Resources and Evaluation*, pages 1625–1628.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

Emi Izumi, Toyomi Saiga, Thepchai Supnithi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2003a. The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques. In *Proc. of the Corpus Linguistics 2003 Conference*, pages 359–366.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003b. Automatic error detection in the Japanese learners' English spoken data. In *Proc. of 41st Annual Meeting of ACL*, pages 145–148.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2005. Error annotation for corpus of Japanese learner English. In *Proc. of 6th International Workshop on Linguistically Annotated Corpora*, pages 71–80.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool, San Rafael.

John Lee and Stephanie Seneff. 2008. Correcting misuse of verb forms. In *Proc. of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology Conference*, pages 174–182.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Proc. of 23rd International Conference on Computational Linguistics, poster volume*, pages 894–900.

Ryo Nagata, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2004. Recognizing article errors based on the three

head words. In *Proc. of Cognition and Exploratory Learning in Digital Age*, pages 184–191.

Ryo Nagata, Takahiro Wakana, Fumito Masui, Atsuo Kawai, and Naoki Isu. 2005. Detecting article errors based on the mass count distinction. In *Proc. of 2nd International Joint Conference on Natural Language Processing*, pages 815–826.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proc. of 44th Annual Meeting of ACL*, pages 241–248.

Katsuaki Okihara. 1985. *English writing (in Japanese)*. Taishukan, Tokyo.

Alla Rozovskaya and Dan Roth. 2010a. Annotating ESL errors: Challenges and rewords. In *Proc. of NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Alla Rozovskaya and Dan Roth. 2010b. Training paradigms for correcting errors in grammar and usage. In *Proc. of 2010 Annual Conference of the North American Chapter of the ACL*, pages 154–162.

Joel Tetreault, Elena Filatova, and Martin Chodorow. 2010a. Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In *Proc. of NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–48.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010b. Using parse features for preposition selection and error detection. In *Proc. of 48nd Annual Meeting of the Association for Computational Linguistics Short Papers*, pages 353–358.

Yukio Tono. 2000. A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora. In *Practical Applications in Language Corpora*, pages 123–132.