Phrase Structure Annotation and Parsing for Learner English

Ryo Nagata
Konan University
8-9-1 Okamoto, Higashinada
Kobe, Hyogo 658-8501, Japan
nagata-acl @hyogo-u.ac.jp.

Keisuke Sakaguchi
Johns Hopkins University
3400 North Charles Street
Baltimore, MD, 21218, USA
keisuke@cs.jhu.edu

Abstract

There has been almost no work on phrase structure annotation and parsing specially designed for learner English despite the fact that they are useful for representing the structural characteristics of learner English. To address this problem, in this paper, we first propose a phrase structure annotation scheme for learner English and annotate two different learner corpora using it. Second, we show their usefulness, reporting on (a) inter-annotator agreement rate, (b) characteristic CFG rules in the corpora, and (c) parsing performance on them. In addition, we explore methods to improve phrase structure parsing for learner English (achieving an F-measure of 0.878). Finally, we release the full annotation guidelines, the annotated data, and the improved parser model for learner English to the public.

1 Introduction

Learner corpora have been essential for NLP tasks related to learner language such as grammatical error correction. They are normally annotated with linguistic properties. In the beginning, attention was mainly focused on grammatical error annotation (Izumi et al., 2004; Díaz-Negrillo et al., 2009; Dale and Kilgarriff, 2011; Ng et al., 2013). Recently, it has been expanded to grammatical annotation — first, Part-Of-Speech (POS) tagging (Díaz-Negrillo et al., 2009; Nagata et al., 2011) and then syntactic annotation (Kepser et al., 2004; Dickinson and Ragheb, 2009; Ragheb and Dickinson, 2012; Ragheb and Dickinson, 2013); syntactic annotation for learner corpora is now intensively studied. Among a variety of studies, a series of work by Ragheb and Dickinson (Dickinson and Ragheb, 2009; Ragheb and Dickinson, 2012; Ragheb and Dickinson, 2013) is important in that they proposed a dependency annotation scheme, theoretically and empirically evaluated it, and revealed its theoretical problems, which gives a good starting point to those who wish to develop a new annotation scheme for learner corpora. Researchers including Foster (2004) and Ott and Ziai (2010) have even started using dependency-annotated learner corpora to develop dependency parsers for learner language.

Although research on syntactic analysis for learner corpora has been making great progress as noted above, it is not yet complete. There are at least three limitations in the previous work: (i) as far as we are aware, there has been almost no work on phrase structure annotation specially designed for learner corpora; (ii) there are no publicly available learner corpora annotated with syntax; (iii) phrase structure parsing performance on learner English has not yet been reported.

The first limitation is that there exists no phrase structure annotation scheme specially designed for learner English. As related work, Foster (2007a; 2007b) and Foster and Andersen (2009) propose a method for creating a pseudo-learner corpus by artificially generating errors in a native corpus with phrase structures. However, the resulting corpus does not capture various error patterns in learner English.

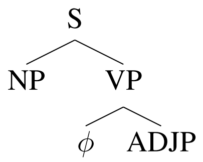
Concerning the second limitation, a corpus greatly increases in value when it is available to the public as has been seen in other domains. Nevertheless, whether dependency or phrase structure, there seems to be no publicly available learner corpora annotated with syntax.

The above two limitations cause the third one that phrase structure parsing performance on leaner English has not yet been reported. For this reason, Cahill (2015) demonstrates how ac-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1837–1847,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

curately an existing parser performs on a pseudo-learner corpus (section 23 of WSJ with errors artificially generated by Foster and Andersen (2009)'s method). Cahill et al. (2014) show the performance of a phrase structure parser augmented by self-training on students' essays, many of which are presumably written by native speakers of English. Tetreault et al. (2010) partially show phrase structure parsing performance concerning preposition usage in learner English, concluding that it is effective in extracting features for preposition error correction. We need to reveal full parsing performance to be able to confirm that this is true for other syntactic categories and whether or not we should use phrase structure parsing to facilitate related tasks such as grammatical error correction and automated essay scoring.

Here, we emphasize that phrase structure annotation has at least two advantages over dependency annotation[1]. First of all, it can directly encode information about word order. This is particularly important because learner corpora often contain errors in word order. For example, phrase structure parsing will reveal in which phrases errors in word order tend to occur as we will partly do in Sect. 3. Second of all, phrase structure rather abstractly represents syntactic information in terms of phrase-to-phrase relations. This means that the characteristics of learner English are represented by means of phrase-to-phrase relations (e.g., context free grammar (CFG) rules) or even as trees. Take as an example, one of the characteristic trees we found in the corpora we have created:



As we will discuss in Sect. 3, this tree suggests the mother tongue interference that the copula is not necessary in adjective predicates in certain languages. It would be linguistically interesting to reveal what CFG rules we need to add to, or subtract from, the native CFG rule set to be able to generate learner English. This is our primary motivation for this work although our other motivations include developing a parser for learner English.

In view of this background, we address the above problems in this paper. Our contributions are three-fold. First, we present a phrase structure annotation scheme for dealing with learner English consistently and reliably. For this, we propose five principles which can be applied to creating a novel annotation scheme for learner corpora. Second, we evaluate the usefulness of the annotation scheme by annotating learner corpora using it. To be precise, we report on inter-annotator agreement rate and characteristic CFG rules in the corpora, and take the first step to revealing phrase structure parsing performance on learner English. In addition, we explore methods to improve phrase structure parsing for learner English. Finally, we release the full annotation guidelines, the annotated corpora, and the improved parser model to the public.

The rest of this paper is structured as follows. Sect. 2 describes the annotation scheme. Sect. 3 explores the annotated learner corpora. Sect. 4 evaluates parsing performance using it.

2 Phrase Structure Annotation Scheme

2.1 General Principles

The annotation scheme is designed to consistently retrieve the structure in the target text that is closest to the writer's intention. The following are the five principles we created to achieve it:

(P1) Consistency-first principle
(P2) Minimal rule set principle
(P3) Locally superficially-oriented principle
(P4) Minimum edit distance principle
(P5) Intuition principle

(P1) states that the most important thing in our annotation scheme is consistency. It is a trade-off between quality and quantity of information; detailed rules that are too complicated make annotation unmanageable yet they may bring out valuable information in learner corpora. Corpus annotation will be useless if it is inconsistent and unreliable no matter how precisely the rules can describe linguistic phenomena. Therefore, this principle favors consistency over completeness. Once we annotate a corpus consistently, we consider adding further detailed information to it.

(P2) also has to do with consistency. The smaller the number of rules is, the easier it becomes to practice the rules. Considering this, if we have several candidates for describing a new linguistic phenomenon particular to learner English, we will choose the one that minimizes the number of modifications to the existing rule set. Note that

[1] We are not arguing that phrase structure annotation is better than dependency annotation; they both have their own advantages, and thus both should be explored.

this applies to the entire rule set; an addition of a rule may change the existing rule set.

(P3) is used to determine the tag of a given token or phrase. As several researchers (Díaz-Negrillo et al., 2009; Dickinson and Ragheb, 2009; Nagata et al., 2011; Ragheb and Dickinson, 2012) point out, there are two ways of performing annotation, according to either superficial (morphological) or contextual (distributional) evidence. For example, in the sentence **My university life is enjoy.*, the word *enjoy* can be interpreted as a verb according to its morphological form or as an adjective (enjoyable) or a noun (enjoyment) according to its context. As the principle itself construes, our annotation scheme favors superficial evidence over distributional. This is because the interpretation of superficial evidence has much less ambiguity and (P3) can determine the tag of a given token by itself as seen in the above example. Distributional information is also partly encoded in our annotation scheme as we discuss in Subsect. 2.2.

(P4) regulates how to reconstruct a correct form of a given sentence containing errors, which helps to determine its phrase structure. The problem is that often one can think of several candidates as possible corrections, which can become a source of inconsistency. (P4) gives a clear solution to this problem. It selects the one that minimizes the edit distance from the original sentence. Note that the edit distances for deletion, addition, and replacement are one, one, and two (deletion and addition), respectively in our definition.

For the cases to which these four principles do not apply, the fifth and final principle (P5) allows annotators to use their intuition. It should be noted, however, that the five principles apply in the above order to avoid unnecessary inconsistency.

2.2 Annotation Rules

Our annotation scheme is based on the POS-tagging and shallow-parsing annotation guidelines for learner English (Nagata et al., 2011), which in turn are based on the Penn Treebank II-style bracketing guidelines (Bies et al., 1995) (which will be referred to as PTB-II, hereafter). This naturally leads us to adopt the PTB-II tag set in ours; an exception is that we exclude the function tags and null elements from our present annotation scheme for annotation efficiency[2]. Accordingly, we revise the above guidelines to be able to describe phrase structures characteristic of learner English.

The difficulties in syntactic annotation of learner English mainly lie in the fact that grammatical errors appear in learner English. Grammatical errors are often classified into three types as in Izumi et al. (2004): omission, insertion, and replacement type errors. In addition, we include other common error types (word order errors and fragments) in the error types to be able to describe learners' characteristics more precisely. The following discuss how to deal with these five error types based on the five principles.

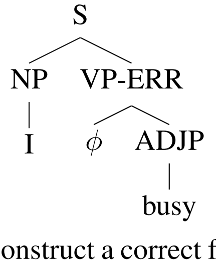
2.2.1 Omission Type Errors

This type of error is an error where a necessary word is missing. For example, some kind of determiner is missing in the sentence **I am student.*

The existing annotation rules in PTB-II can handle most omission type errors. For instance, the PTB-II rule set would parse the above example as "(S (NP I) (VP am (NP student).))." Note that syntactic tags for irrelevant parts are omitted in this example (and hereafter).

A missing head word may be more problematic. Take as an example the sentence **I busy.* where a verb is missing. The omission prevents the rule S → NP VP from applying to it. If we created a new rule for every head-omission with no limitation, it would undesirably increase the number of rules, which violates (P2).

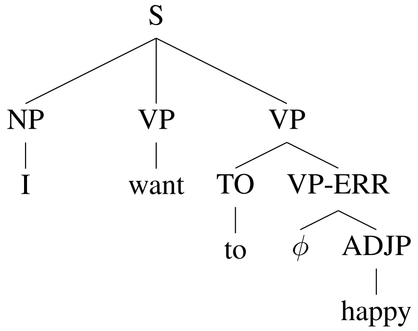
To handle head-omissions, we propose a function tag *-ERR*. It denotes that a head is missing in the phrase in question. The function tag makes it possible to apply the PTB-II rule set to sentences containing head-omissions as in:



We need to reconstruct a correct form of a given sentence to determine whether or not a head word is missing. We use Principle (P4) for solving the problem as discussed in Sect. 2.1. For instance, the sentence **I want to happy.* can be corrected as either *I want to be happy.* (edit distance is one; an addition of a word) or *I want happiness.* (three; two deletions and an addition). Following (P4), we select the first correction that minimizes the edit

[2] We will most likely include them in a future version.

distance, resulting in:



2.2.2 Insertion Type Errors

An insertion type error is an error where an extra word is used incorrectly. For example, the word *about* is an extra word in the sentence **She discussed about it.*

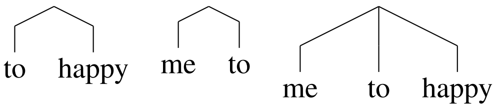
Insertion type errors are more problematic than omission type errors. It is not trivial how to annotate an erroneous extra word. On the one hand, one can argue that the extra word *about* is a preposition from its morphological form. On the other hand, one can also argue that it is not, because the verb *discuss* takes no preposition. As with this example, insertion type errors involve an ambiguity between superficial and distributional categories.

Principles (P2) and (P3) together solve the ambiguity. According to (P3), one should always stick to the superficial evidence. For example, the extra word *about* should be tagged as a preposition. After this, PTB-II applies to the rest of the sentence, which satisfies (P2). As a result, one would obtain the parse "(S (NP She) (VP discussed (PP (IN about) (NP it)))).)."

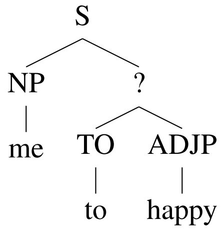
Insertion type errors pose a more vital problem in some cases. Take as an example the sentence **It makes me to happy.* where the word *to* is erroneous. As before, one can rather straightforwardly tag it as a preposition, giving the POS sequence:

**It/PRP makes/VBZ me/PRP to/TO happy/JJ ./.*

However, none of the PTB-II rules applies to the POS sequence *TO JJ* to make a phrase. This means that we have to create a new rule for such cases. There are at most three possibilities of grouping the words in question to make a phrase:



Intuitively, the first one seems to be the most acceptable. To be precise, the second one assumes a postposition, contrary to the English preposition system. The third one assumes a whole new rule generating a phrase from a personal pronoun, a preposition, and an adjective into a phrase. Thus, they cause significant modifications to PTB-II, which violates (P2). In contrast, a preposition normally constitutes a prepositional phrase with another phrase (although not normally with an adjective phrase). Moreover, the first grouping would produce for the rest of the words the perfect phrase structure corresponding to the correct sentence without the preposition *to*:



which satisfies (P2) unlike the second and third ones. Accordingly, we select the first one.

All we have to do now is to name the phrase *to happy*. There is an ambiguity between PP and ADJP, both of which can introduce the parent S. The fact that a preposition constitutes a prepositional phrase with another phrase leads us to select PP for the phrase. Furthermore, the tag of a phrase is normally determined by the POS of one of the immediate constituents, if any, that is entitled to be a head (i.e., the headedness). Considering this, we select PP in this case, which would give the parse to the entire sentence as follows:"(S (NP It) (VP makes (S (NP me) (PP (TO to) (ADJP happy)))).)."

In summary, for insertion errors to which PTB-II do not apply, we determine their phrase structures as follows: (i) intuitively group words into a phrase, minimizing the number of new rules added (it is often helpful to examine whether an existing rule is partially applicable to the words in question); (ii) name the resulting phrase by the POS of one of the immediate children that is entitled to be a head.

2.2.3 Replacement Type Errors

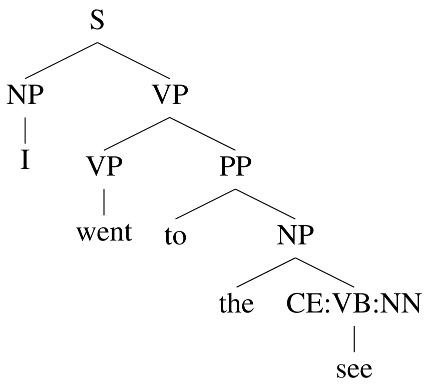
A replacement type error is an error where a word should be replaced with another word. For example, in the sentence: **I often study English conversation.*, the verb *study* should be replaced with a more appropriate verb such as *practice*.

To handle replacement type errors systematically, we introduce a concept called *POS class*, which is a grouping of POS categories defined as

Class	Members
Noun	NN, NNS, NNP, NNPS
Verb	VB, VBP, VBZ, VBD
Adjective	JJ, JJR, JJS
Adverb	RB, RBR, RBS
Participle	VBN, VBG

Table 1: POS class.

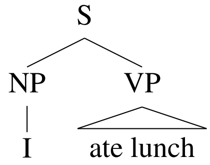
in Table 1; POS tags that are not shown in Table 1 form a POS class by itself. If the replacement in question is within the same POS class, it is annotated following Principles (P2) and (P3). Namely, the erroneous word is tagged according to its superficial form and the rest of the sentence is annotated by the original rule set, which avoids creating new rules[3]. If the replacement in question is from one POS class to another, we will need to take special care because of the ambiguity between superficial and distributional POS categories. For example, consider the sentence **I went to the see.* where the word *see* is used as a noun, which is not allowed in the standard English, and the intention of the learner is likely to be *sea* (from the surrounding context). Thus, the word *see* is ambiguous between a verb and a noun in the sentence. To avoid the ambiguity, we adopt a two layer-annotation scheme (Díaz-Negrillo et al., 2009; Nagata et al., 2011; Ragheb and Dickinson, 2012) to include both POSs. In our annotation scheme, we use a special tag (CE) for the replacement error and encode the two POSs as its attribute values as in *CE:VB:NN*. Then we can use the distributional POS tag to annotate the rest of the sentence. For example, the above example sentence would give a tree:



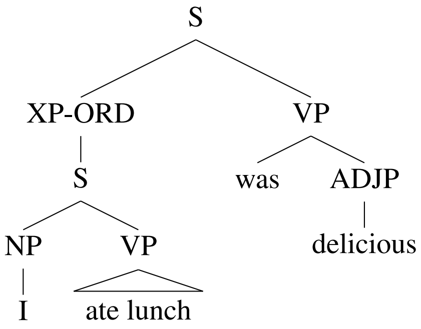
2.2.4 Errors in Word Order

Errors in word order often appear in learner English. A typical example would be the reverse of the subject-object order: **This place like my friends.* (correctly, *My friends like this place.*).

Principles (P2) and (P3) again play an important role in handling errors in word order. We first determine the POS tag of each word according to its morphological form. This is rather straightforward because errors in word order do not affect the morphological form. Then we determine the whole structure based on the resulting POS tags, following Principle (P2); if rules in PTB-II apply to the sentence in question, we parse it according to them just as in the above example sentence: "(S (NP This place) (VP like (NP my friends)).)" Even if any of the existing rules do not apply to a part of the sequence of the given POS tags, we stick to Principle (P3) as much as possible. In other words, we determine partial phrase structures according to the given POS sequence to which the existing rule set applies. Then we use the XP-ORD tag to put them together into a phrase. As an example, consider the sentence **I ate lunch was delicious.* (correctly, *The lunch I ate was delicious.*). According to the superficial forms and local contexts, the phrase *I ate lunch* would form an S:



However, the relations of the S to the rest of the constituents are not clear. Here, we use the XP-ORD tag to combine the S with the rest together:



2.2.5 Fragments

In learner corpora, sentences are sometimes incomplete. They are called fragments (e.g., missing main clause: *Because I like it.*).

Fortunately, there exists already a tag for fragments in PTB-II: FRAG. Accordingly, we use

[3] This means that spelling and morphological errors are not directly coded in our annotation scheme as in *He/PRP has/VBZ a/DT <u>books/NNS</u>.*

it in our annotation scheme as well. For example, the above example would give the parse "(FRAG (SBAR Because (S (NP I (VP like (NP it))))).)" An exception is incomplete sentences which are defined as S in the bracketing guidelines for biomedical texts (Warner et al., 2012). We tag such incomplete sentences as S following the convention. For example, an adjective phrase can form an S (e.g., (S (ADVP Beautiful)!)).

2.2.6 Unknown Words and Phrases

There are cases where one cannot tell the tag of a given word. We use the UK tag for such words (e.g., *Everyone is don/UK)*.

Even if its tag is unknown, it is somehow clear in some cases that the unknown word is the head word of the phrase just as in the above example. In that case, we use the UP tag so that it satisfies the rule about the headedness of a phrase we have introduced in Subsect. 2.2.2. Based on this, the above example would give the parse "(S (NP everyone) (VP is (UP (UK don))).)"

For a phrase whose head word is unknown due to some error(s) in it, we use the XP tag instead of the UP tag. As a special case of XP, we use the XP-ORD tag to denote the information that we cannot determine the head of the phrase because of an error in word order.

3 Corpus Annotation

We selected the Konan-JIEM (KJ) learner corpus (Nagata et al., 2011) (beginning to intermediate levels) as our target data. It is manually annotated with POSs, chunks, and grammatical errors, which helps annotators to select correct tags. We also included in the target data a part of the essays in ICNALE (Ishikawa, 2011) consisting of a variety of learners (beginning to advanced levels[4]) in Asia (China, Indonesia, Japan, Korea, Taiwan, Thailand, Hong Kong, Singapore, Pakistan, Philippines). Table 2 shows the statistics on the two learner corpora.

Two professional annotators[5] participated in the annotation process. One of them first annotated the KJ data and double-checked the results. Between the first and second checks, we discussed the results with the annotator. We revised the annotation scheme based on the discussion which resulted in the present version. Then the second annotator annotated a part of the KJ data to evaluate the consistency between the two annotators. We took out 11 texts (955 tokens) as a development set. The second annotator annotated it using the revised annotation scheme where she consulted the first annotator if necessary. After this, we provided her with the differences between the results of the two annotators. Finally, the first annotator annotated the data in ICNALE while the second independently another part of the KJ data and a part of the ICNALE data (59 texts, 12,052 tokens in total), which were treated as a test set.

Table 3 shows inter-annotator agreement measured in recall, precision, F-measure, complete match rate, and chance-corrected measure (Skjærholt, 2014). We used the EVALB tool[6] with the Collins (1997)'s evaluation parameter where we regarded the annotation results of the first annotator as the gold standard set. We also used the syn-agreement tool[7] to calculate chance-corrected measure. It turns out that the agreement is very high. Even in the test set, they achieve an F-measure of 0.928 and a chance-corrected measure of 0.982. This shows that our annotation scheme enabled the annotators to consistently recognize the phrase structures in the learner corpora in which grammatical errors frequently appear. The comparison between the results of the two annotators shows the major sources of the disagreements. One of them is annotation concerning adverbial phrases. In PTB-II, an adverbial phrase between the subject NP and the main verb is allowed to be a constituent of the VP (e.g., (S (NP I) (VP (ADVP often) go))) and also of the S (e.g., (S (NP I) (ADVP often) (VP go))). Another major source is the tag FRAG (fragments); the annotators disagreed on distinguishing between FRAG and S in some cases.

The high agreement shows that the annotation scheme provides an effective way of consistently annotating learner corpora with phrase structures. However, one might argue that the annotation does not represent the characteristics of learner English well because it favors consistency (and rather simple annotation rules) over completeness.

To see if the annotation results represent the

[4] The details about the proficiency levels are available in `http://language.sakura.ne.jp/icnale/about.html`

[5] The annotators, whose mother tongue is Japanese, have a good command of English. They have engaged in corpus annotation including phrase structure annotation for around 20 years.

[6] `http://nlp.cs.nyu.edu/evalb/`

[7] `https://github.com/arnsholt/syn-agreement`

Corpus	# essays	# sentences	# tokens	# errors/token	# errors/sentence
KJ	233	3,260	30,517	0.15	1.4
ICNALE	134	1,930	33,913	0.08	1.4

Table 2: Statistics on annotated learner corpora.

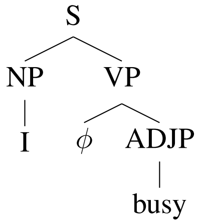
Set	R	P	F	CMR	CCM
Development	0.981	0.981	0.981	0.913	0.995
Test	0.919	0.927	0.928	0.549	0.982

Table 3: Inter-annotator agreement measured in Recall (R), Precision (P), F-measure (F), Complete Match Rate (CMR), and Chance-Corrected Measure (CCM).

characteristics of learner English, we extracted characteristic CFG rules from them. The basic idea is that we compare the CFG rules obtained from them with those from a native corpus (the Penn Treebank-II)[8]; we select as characteristic CFG rules those that often appear in the learner corpora and not in the native corpus. To formalize the extraction procedures, we denote a CFG rule and its conditional probability as $A \rightarrow B$ and $p(B|A)$, respectively. Then we define the score for $A \rightarrow B$ by $s(A \rightarrow B) = \log \frac{p_L(B|A)}{p_N(B|A)}$ where we distinguish between learner and native corpora by the subscripts L and N, respectively. We estimate $p(B|A)$ by expected likelihood estimation. Note that we remove the function tags to reduce the differences in the syntactic tags in both corpora when we calculate the score.

Table 4 shows the top 10 characteristic CFG rules sorted in descending and ascending order according to their scores, which correspond to overused and underused rules in the learner corpora, respectively. Note that Table 4 excludes rules consisting of only terminal and/or preterminal symbols to focus on the structural characteristics. Also, it excludes rules containing a Quantifier Phrase (QP; e.g., (NP (QP 100 million) dollars)), which frequently appear and is one of the characteristics in the native corpus.

In the overused column, CFG rules often contain the ϕ element. At first sight, this does not seem so surprising because ϕ never appears in the native corpus. However, the rules actually show in which syntactic environment missing heads tend to occur. For example, the CFG rule *PP* $\rightarrow$ ϕ *S* shows that prepositions tend to be missing in the prepositional phrase governing an S as in **I am good _ doing this*, which we had not realized before this investigation. More interestingly, the CFG rule *VP* $\rightarrow$ ϕ *ADJP* reveals that an adjective phrase can form a verb phrase without a verb in learner English. Looking into the annotated data shows that the copula is missing in predicative adjectives as in the tree:



This suggests the transfer of the linguistic system that the copula is not necessary or may be omitted in predicate adjectives in certain languages such as Japanese and Chinese. Similarly, the rule *VP* $\rightarrow$ ϕ *NP* shows in which environment a verb taking the object tends to be missing. Out of the 28 instances, 18 (64%) are in a subordinate clause, which implies that learners tend to omit a verb when more than one verb appear in a sentence.

The second rule *S* $\rightarrow$ *XP VP .* implies that the subject NP cannot be recognized because of a combination of grammatical errors (c.f., *S* $\rightarrow$ *NP VP .*). The corpus data show that 21% of *XP* in *S* $\rightarrow$ *XP VP .* are actually *XP-ORD* concerning an error in a relative clause just as shown in the tree in Subsect. 2.2.4. Some of the learners apparently have problems in appropriately using relative clauses in the subject position. It seems that the structure of the relative clause containing another verb before the main verb confuses them.

Most of the underused CFG rules are those that introduce rather complex structures. For exam-

[8] To confirm that the extracted characteristics are not influenced by the differences in the domains of the two corpora, we also compared the learner data with the native speaker sub-corpus in ICNALE that is in the same domain. It turned out that the extracted CFG rules, were very similar to those shown in Table 4.

Overuse	Score	Underuse	Score
PP → ϕ NP	9.0	NP → NP , NP ,	-4.6
S → XP VP .	7.2	S → NP NP	-2.7
PP → IN IN S	6.7	S → NP VP . ”	-2.6
S → XP .	6.6	ADVP → NP RBR	-2.5
VP → ϕ ADJP	6.5	S → S , NP VP .	-2.4
VP → ϕ NP	6.3	NP → NP , SBAR	-2.4
SBAR → IN NN TO S	6.1	SBAR → WHPP S	-2.3
PP → ϕ S	6.1	VP → VBD SBAR	-2.2
S → ADVP NP ADVP VP .	5.8	S → NP PRN VP .	-2.2
PP → IN TO NP	5.7	S → PP , NP VP . ”	-2.1

Table 4: Characteristic CFG rules.

ple, the eighth rule *VP → VBD SBAR* implies a structure such as *He thought that* $\cdots$. The underused CFG rules are a piece of the evidence that this population of learners of English cannot use such complex structures as fluently as native speakers do. Considering this, it will be useful feedback to provide them with the rules (transformed into interpretable forms). As in this example, phrase structure annotation should be useful not only for second language acquisition research but also for language learning assistance.

4 Parsing Performance Evaluation

We tested the following two state-of-the-art parsers on the annotated data: Stanford Statistical Natural Language Parser (ver.2.0.3) (de Marneffe et al., 2006) and Charniak-Johnson parser (Charniak and Johnson, 2005). We gave the tokenized sentences to them as their inputs. We used again the EVALB tool with the Collins (1997)'s evaluation parameter.

Table 5 shows the results. To our surprise, both parsers perform very well on the learner corpora despite the fact that it contains a number of grammatical errors and also syntactic tags that are not defined in PTB-II. Their performance is comparable to, or even better than, that on the Penn Treebank (reported in Petrov (2010)).

To achieve further improvement, we augmented the Charniak-Johnson parser with the learner data. We first retrained its parser model using the 2-21 sections of Penn Treebank Wall Street Journal (hereafter, WSJ) as training data and its 24 section as development data, following the settings shown in Charniak and Johnson (2005). We then added the learner corpora to the training data using six-fold cross validation. We split it into six parts, each of which approximately consisted of 61 essays, used one sixth as test data, another one sixth as development data instead of the 24 section, and retrained the parser model using the development data and the training data consisting of the remaining four-sixths part of the learner data and the 2-21 sections of WSJ. We also conducted experiments where we copied the four sixths of the learner data n times ($1 \leq n \leq 50$) and added them to the training data to increase its weight in retraining.

Parser	R	P	F	CMR
Stanford	0.812	0.832	0.822	0.398
Charniak-Johnson	0.845	0.865	0.855	0.465

Table 5: Parsing performance on learner English.

Figure 1 shows the results. The simple addition of the learner data ($n = 1$) already outperforms the parser trained only on the 2-21 sections of WSJ ($n = 0$) in both recall and precision, achieving an F-measure of 0.866 and a complete match rate of 0.515. The augmented parser model particularly works well on recognizing erroneous fragments in the learner data; F-measure improved to 0.796 ($n = 1$) from 0.683 ($n = 0$) in the sentences containing fragments (i.e., FRAG) (46 out of the 111 sentences that were originally erroneously parsed made even a complete match). It was also robust against spelling errors. The performance further improves as the weight n increases (up to $F = 0.878$ when $n = 24$), which shows the effectiveness of using learner corpus data as training data.

Figure 2 shows the parsing performance of the Charniak-Johnson parser in each sub-corpus of

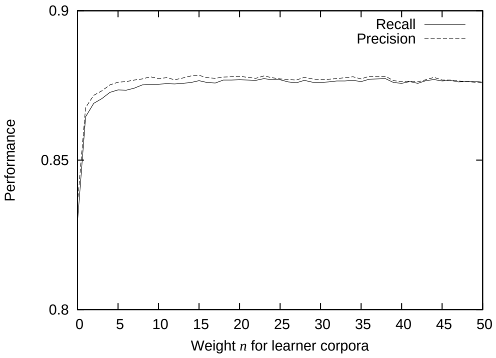


Figure 1: Relation between learner corpus size in training data and parsing performance.

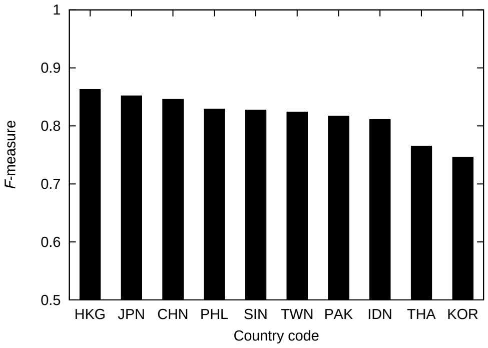


Figure 2: Parsing performance in each sub-corpus.

ICNALE (classified by country code[9]). In most of the sub-corpora, the parser achieves an F-measure of 0.800 or better. By contrast, it performs much worse on the Korean sub-corpus. The major reason for this is that it contains a number of word order errors (i.e., XP-ORD); to be precise, 27 instances compared to zero to two instances in the other sub-corpora. Similarly, FRAG is a source of parsing errors in the Thai sub-corpus. We need further investigation to determine whether the differences in parsing performance are due to the writers' mother tongue or other factors (e.g., proficiency).

We can summarize the findings as follows: (1) the state-of-the-art phrase structure parsers for native English are effective even in parsing learner English; (2) they are successfully augmented by learner corpus data; (3) the evaluation results support the previous report (Tetreault et al., 2010) that they are effective in extracting parse features for grammatical error correction (and probably for related NLP tasks such as automated essay scoring); (4) however, performance may vary depending on the writer's mother tongue and/or other factors, which we need further investigation to confirm.

5 Conclusions

This paper explored phrase structure annotation and parsing specially designed for learner English. Sect. 3 showed the usefulness of our phrase structure annotation scheme and the learner corpora annotated using it. The annotation results exhibited high consistency. They also shed light on (at least, part of) the characteristics of the learners of English. Sect. 4 further reported on the performance of the two state-of-the-art parsers on the annotated corpus, suggesting that they are accurate for providing NLP applications with phrase structures in learner English. All these findings support the effectiveness of our phrase structure annotation scheme for learner English. It would be much more difficult to conduct similar analyses and investigations without the phrase structure annotation scheme and a learner corpus annotated based on it. The annotation guidelines, the annotated data, and the parsing model for learner English created in this work are now available to the public[10].

In our future work, we will evaluate parsing performance on other learner corpora such as ICLE (Granger et al., 2009) consisting of a wide variety of learner Englishes. We will also extend phrase structure annotation, especially working on function tags.

Acknowledgments

We would like to thank Shin'ichiro Ishikawa, who created the ICNALE corpus, for providing us with the data and Arne Skærholt for the assistance to run the syn-agreement tool. We would also like to thank the three anonymous reviewers for their valuable feedback. This work was partly supported by Grant-in-Aid for Young Scientists (B) Grant Number JP26750091.

[9] Ideally, it would be better to use sub-corpora classified by their mother tongues. Unfortunately, however, only country codes are provided in ICNALE.

[10] We released the Konan-JIEM corpus with phrase structures on March 2015, which is available at `http://www.gsk.or.jp/en/catalog/gsk2015-a/`. We annotated the existing ICNALE, which was created by Dr. Ishikawa and his colleagues, with phrase structures. We released the data on Jun 2016, which is available at `http://language.sakura.ne.jp/icnale/`

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II-style Penn treebank project.

Aoife Cahill, Binod Gyawali, and James V. Bruno. 2014. Self-training for parsing learner text. In *Proc. of 1st Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 66–73.

Aoife Cahill. 2015. Parsing learner text: to Shoehorn or not to Shoehorn. In *Proc. of 9th Linguistic Annotation Workshop*, pages 144–147.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine N-best parsing and MaxEnt discriminative reranking. In *Proc. of 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proc. of 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proc. of 13th European Workshop on Natural Language Generation*, pages 242–249.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. of 5th International Conference on Language Resources and Evaluation*, pages 449–445.

Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2009. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154.

Markus Dickinson and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In *Proc. of 8th Workshop on Treebanks and Linguistic Theories*, pages 59–70.

Jennifer Foster and Øistein E. Andersen. 2009. GenERRate: Generating errors for use in grammatical error detection. In *Proc. of 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90.

Jennifer Foster. 2004. Parsing ungrammatical input: An evaluation procedure. In *Proc. of 4th International Conference on Language Resources and Evaluation*, pages 2039–2042.

Jennifer Foster. 2007a. Treebanks gone bad: generating a treebank of ungrammatical English. In *2007 Workshop on Analytics for Noisy Unstructured Data*, pages 39–46, Jan.

Jennifer Foster. 2007b. Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition*, 10(3):129–145, Dec.

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

Shinichiro Ishikawa, 2011. *A new horizon in learner corpus studies: The aim of the ICNALE project*, pages 3–11. University of Strathclyde Publishing, Glasgow.

Emi Izumi, Toyomi Saiga, Thepchai Supnithi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, 12(2):119–125.

Stephan Kepser, Ilona Steiner, and Wolfgang Sternefeld. 2004. Annotating and querying a treebank of suboptimal structures. In *Proc. of 3rd Workshop on Treebanks and Linguistic Theories*, pages 63–74.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proc. 17th Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

Niels Ott and Ramon Ziai. 2010. Evaluating dependency parsing performance on German learner language. In *Proc. of 9th International Workshop on Treebanks and Linguistic Theories*, pages 175–186.

Slav Petrov. 2010. Products of random latent variable grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proc. of 24th International Conference on Computational Linguistics*, pages 965–974.

Marwa Ragheb and Markus Dickinson. 2013. Inter-annotator agreement for dependency annotation of learner language. In *Proc. of 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Arne Skjærholt. 2014. A chance-corrected measure of inter-annotator agreement for syntax. In *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics*, pages 934–944.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. 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.

Colin Warner, Arrick Lanfranchi, Tim O'Gorman, Amanda Howard, Kevin Gould, and Michael Regan. 2012. Bracketing biomedical text: An addendum to Penn Treebank II guidelines.