Phrase Structure Annotation and Parsing for Learner English

Keisuke Sakaguchi[†] and Ryo Nagata[††]

Learner English often contains grammatical errors with structural characteristics such as omissions, insertions, substitutions, and word order errors. These errors are not covered by the existing context-free grammar (CFG) rules. Therefore, it is not at all straightforward how to annotate learner English with phrase structures. Because of this limitation, there has been almost no work on phrase structure annotation for learner corpora despite its importance and usefulness. To address this issue, we propose a phrase structure annotation scheme for learner English, that consists of five principles. We apply the annotation scheme to two different learner corpora and show (i) its effectiveness at consistently annotating learner English with phrase structure (i.e., high inter-annotator agreement); (ii) the structural characteristics (CFG rules) of learner English obtained from the annotated corpora; and (iii) phrase structure parsing performance on learner English for the first time. We also release the annotation guidelines, the annotated data, and the parser model to the public.

Key Words: *Learner English, Phrase Structure, Corpus Annotation, Parsing*

1 Introduction

Learner English contains various types of noise such as grammatical errors that deteriorate the performance of standard Natural Language Processing (NLP) tools (e.g., tokenizers and parsers) trained on canonical texts. Therefore, it is generally more difficult to analyze learner corpora than native corpora.

In the NLP community, grammatical error annotations have been mainly focused on the development of systems that correct grammatical errors (Izumi, Saiga, Supnithi, Uchimoto, and Isahara 2004; Díaz-Negrillo, Meurers, Valera, and Wunsch 2009; Dale and Kilgarriff 2011; Ng, Wu, Wu, Hadiwinoto, and Tetreault 2013). Some recent work focuses more on linguistic annotation. For example, Díaz-Negrillo et al. (2009) and Nagata, Whittaker, and Sheinman (2011) annotated a learner corpus with part-of-speech (POS) labels. Even syntactic annotation for learner corpora is now intensively studied. As related work, Foster (2007a, 2007b) and Foster and Andersen (2009) proposed a method for creating a pseudo-learner corpus by artificially generating errors in a parsed native corpus. A series of studies by Ragheb and Dickinson (2009, 2012, 2013) is

[†] Johns Hopkins University
[††] Konan University

important in that they proposed a dependency annotation scheme, theoretically and empirically evaluated it, and revealed its theoretical problems, creating a good starting point for those who wish to develop a new annotation scheme for learner corpora. Recently, Berzak, Kenney, Spadine, Wang, Lam, Mori, Garza, and Katz (2016) proposed a universal dependency annotation scheme for learner English and released a learner corpus annotated based on it.

Although syntactic annotation for learner English has been received attention, one important class is still missing, that is, *phrase structure annotation*. One of our motivations for developing a phrase structure annotation for learner English is the fact that no one has created such dataset and it has not been easy to analyze phrase structures in learner English.[1] The dependency annotation proposed by Berzak et al. (2016) does not have a clear guideline or principle for annotating missing head words and word order errors, because their scheme does not allow a word/symbol to be inserted in the learners' text. In addition, linguistic analyses for language acquisition research generally employs phrase structures rather than dependencies (Kimura 2003; Narita 2013).

Phrase structure has at least two advantages over dependency. First, phrase structures can represent more abstract notions (i.e., phrases) explicitly, whereas dependency structures aim to extract relationships between words. Of course, it is possible to extract phrase-level information from dependency structures, but it requires extra work. Second, phrase structures can deal with word orders directly. For example, "*I ate the lunch was delicious. (The lunch I ate was delicious.)" has a word order error. Phrase structures can deal with the word order error by simply swapping the phrases NP_1 and $SBAR_2$ in Figure 1, whereas dependency structures have

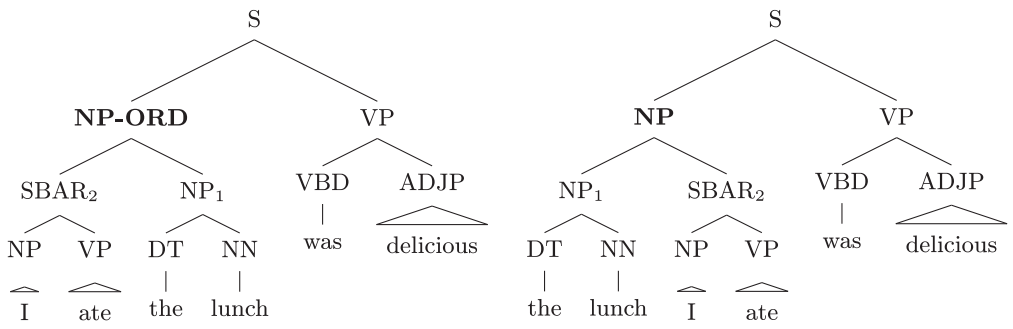


Fig. 1 Phrase structure representation for a word order error example (left: original sentence, right: corrected sentence)

[1] Manual analysis is time and cost inefficient, and standard parsers (trained on native English corpora) is not designed for parsing learner English.

difficulty describing word order errors (especially when more than two words are involved) and/or require extra labels that represent the word order errors (Figure 2). It is also difficult to define the range of word order errors in dependency annotation. As shown in the blocks in Figure 2 (third line), the dependency annotation lacks the explicit subtree spans (*I ate* and *the lunch*) needed to recover the correct word order. Regarding the difference between phrase structure and dependency, we would like to emphasize that the primary goal of this paper is to present the first phrase structure annotation scheme for learner English. We are not arguing that phrase structure annotation is better than dependency annotation in all respects; they both have their own advantages, and thus both should be explored.

At the same time, it is not clear at all how to annotate learner corpora with phrase structures, because they contain various grammatical errors including omissions, insertions, substitutions, and word order errors as mentioned above. These specific errors hinder the application of conventional context-free grammar (CFG) rules to learner corpora. For example, it is not straightforward to annotate a sentence with the head word omission error as in *I *ϕ/am busy.* It is not hard to

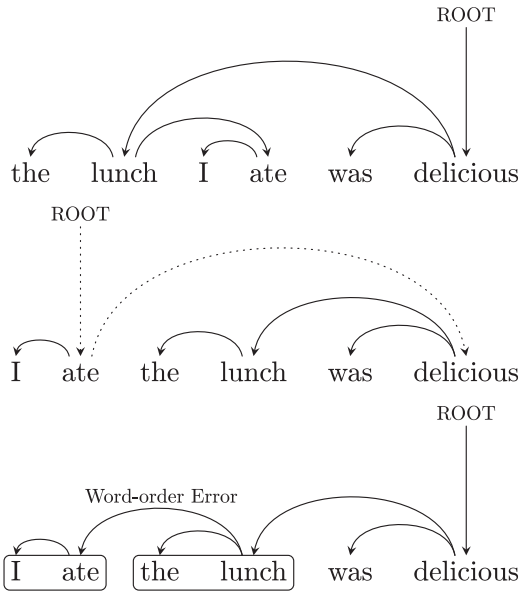
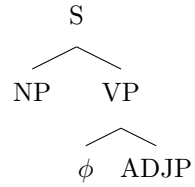


Fig. 2 Universal dependency structure representation for a word order error example. The sentence at the top shows the reference (correct sentence) with oracle dependency. The middle sentence contains a word order error with the automated parsing result of the Stanford Parser. Incorrect arcs are indicated by dotted lines. The bottom sentence shows the universal dependency annotation for the word order error. The spans and subtrees are grouped for clarity. Note that the universal dependency structure does not provide the span information.

imagine that many questions immediately arise: what is the head word of *busy*, and can we define a phrase (non-terminal symbol) that connects *I* and *busy*? For a sentence with a word insertion error such as "*I am *a busy*", there are no CFG rules that can explain the ungrammatical phrase *a busy*. We may introduce a number of new CFG rules (e.g., S → NP ADJP, and VP → DET ADJP) to cover these ungrammatical structures. However, these ad hoc additions increase the number of CFG rules unnecessarily, which in turn lead to a lack of *consistency* in phrase structure annotation. In other words, without the careful design of a phrase structure annotation scheme, annotators would hardly agree on which rule to use from among the various new rules.

In light of all of this, the fundamental question is the following: Is it possible to create an annotation scheme for learner English that achieves consistent phrase structure annotation, and if it is, how? To address these questions, we propose a phrase structure annotation scheme for learner corpora with five principles (Section 2), that achieves consistent and reliable phrase structure annotation. Based on it, we create annotation rules to handle noise in learner English. In the evaluation, we show a reasonably high inter-annotator agreement rate (i.e., significant consistency) for our annotation scheme on two different learner corpora (Section 3). We also present detailed analyses of characteristic phrase structures. For example, we show a structural representation, that likely corresponds to mother tongue interference, where the copula is missing in adjective predicates (in Japanese for example), as in:



Finally, we report on phrase structure parsing performance on learner English for the first time. We also show that a model trained on it can improve parsing performance (0.878 in *F*-measure) in Section 4. Note that there have been a very limited number of publicly available learner corpora that are syntactically annotated; to the best of our knowledge, Berzak et al. (2016) recently released a publicly available learner corpus annotated with dependency (but not phrase structure). This inhibits effective investigations and applications using learner language corpora such as automated grammatical error correction, automated scoring and native language identification. Thus, we release the full annotation guidelines, annotated corpora and improved parser model to the public.

To summarize, our contributions in this paper are as follows:

- We propose a phrase structure annotation scheme and annotation rules that deal with noisy learner English.
- We use them to actually annotate two learner corpora with phrase structures.
- We empirically show that they are effective at consistent phrase structure annotation for learner English.
- We reveal characteristic phrase structures extracted from the annotated learner corpora.
- We report phrase structure parsing performance on learner English for the first time.
- We release the annotation guidelines, annotated corpora, and parser model.

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 this goal:

(P1) Consistency-first
(P2) Minimal rule set
(P3) Superficially-oriented on a local level
(P4) Minimum edit distance
(P5) Intuition

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

(P2) also has to do with consistency. A smaller number of rules makes it easier to practice them. 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 this applies to the entire rule set; the 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 an 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 evidence. 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 Section 2.2.

(P4) regulates how to reconstruct the correct form of a given sentence containing errors, which helps to determine its phrase structure. The problem is that one can often 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 option that minimizes the word 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.[2]

For the cases to which these four principles do not apply, the fifth and final principle (P5) allows annotators to use their intuition. Note, however, that the five principles are applied 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, Ferguson, Katz, and MacIntyre 1995) (referred to as PTB-II, hereafter). This naturally leads us to adopt the PTB-II tag set in our scheme; an exception is that we exclude the function tags and null elements from our present annotation scheme for annotation efficiency.[3] Accordingly, we revise the above guidelines to be able to describe phrase structures characteristic of learner English.

The difficulties in the 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) to describe learners'

[2] This definition allows us to distinguish replacement from deletion and addition in terms of edit distance, which in turn often helps us select a unique interpretation from several candidates as is shown in later sections.

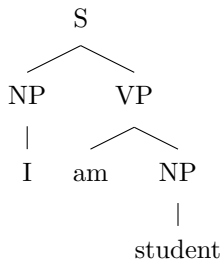
[3] We will most likely include them in a future version.

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.*

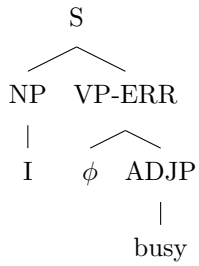
The existing annotation rules in PTB-II can handle most omission type errors. For instance, the PTB-II rule set parses the above example as follows.



Note that syntactic tags for the irrelevant parts are omitted in this example (and hereafter).

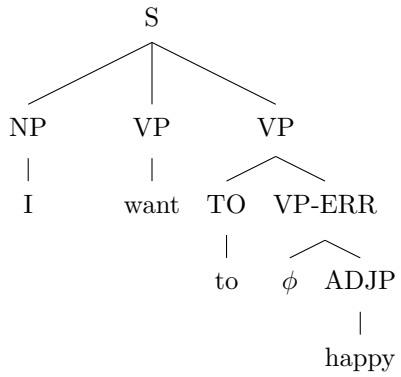
A missing head word may be more problematic. Take as an example the sentence "**I busy.*" where a verb is missing. This omission prevents the rule S → NP VP from being applicable in this case. If we created a new rule for every head-omission with no limitation, it would undesirably increase the number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

To handle head-omissions, we propose the 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 the correct form of a given sentence to determine whether or not a head word is missing. We use (P4) for solving the problem as discussed in Section 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.* (the edit distance three; two deletions

and an addition). Following (P4), we select the first correction that minimizes the edit distance, which results in the following schematic representation:

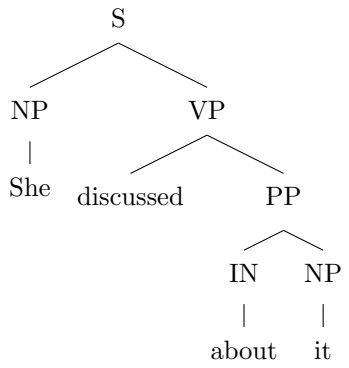


2.2.2 Insertion Type Errors

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 to determine 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.

(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 is applied to the rest of the sentence, which satisfies (P2). As a result, one obtains the parse

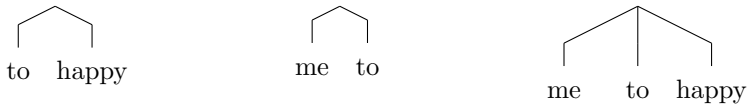


Insertion type errors pose a more vital problem in some cases. Take as an example the

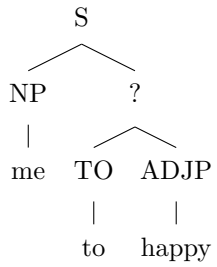
sentence **It makes me to happy.* where the word *to* is erroneous. As before, one can rather straightforwardly tag it as a preposition, giving the POS sequence:

**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:

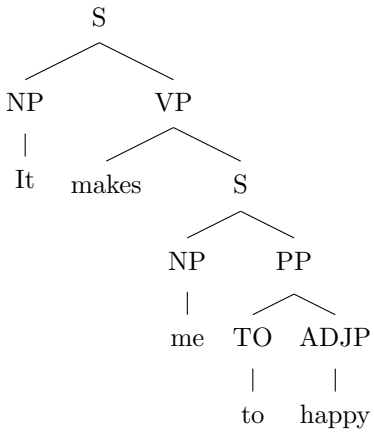


Intuitively, the first one seems to be the most acceptable. To be precise, the second one assumes a postposition, contrary to the English preposition system. The third one assumes a whole new rule generating a phrase from a personal pronoun, a preposition, and an adjective into a phrase. Thus, they cause 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 produces for the rest of the words the perfect phrase structure corresponding to the correct sentence without the preposition *to*:



which satisfies (P2), unlike the last two grouping possibilities listed above. 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 give the parse for the entire sentence as follows:



In summary, for insertion errors to which PTB-II does 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); and (ii) name the resulting phrase using the POS of one of the immediate children that is entitled to be a head.

2.2.3 Replacement Type Errors

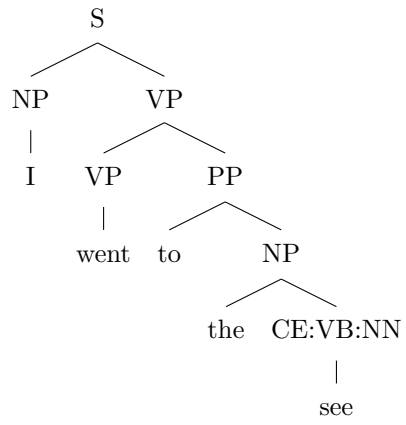
A replacement type error is an error where a word should be replaced with another word. For example, in the sentence "**I often study English conversation.*", the verb *study* should be replaced with a more appropriate verb such as *practice*.

To handle replacement type errors systematically, we introduce a concept called *POS class*, which is a grouping of POS categories as defined in Table 1; POS tags that are not shown in Table 1 form a separate POS class. If the replacement in question is within the same POS class, it is annotated following (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

Table 1 POS class

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

creating new rules.[4] If the replacement in question is from one POS class to another, we 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 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 this 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: cognitive error) for the replacement error and encode the two POSs as its attribute values as in *CE:VB:NN*. We can then use the distributional POS tag to annotate the rest of the sentence. For example, the above example sentence give the tree:



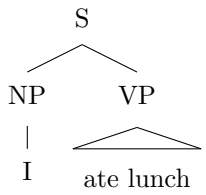
2.2.4 Errors in Word Order

Errors in word order often appear in learner English. Typical word order errors are noun-adjective phrases in Romance languages and the reverse of the subject-object order in elementary-level Japanese learners: **This place like my friends.* (*My friends like this place.*).

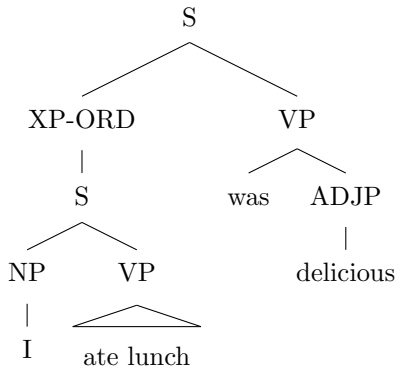
(P2) and (P3) again play an important role in handling errors in word order. We first determine the POS tag of each word according to its morphological form. This is rather straightforward because errors in word order do not affect the morphological form. Then we determine the whole structure based on the resulting POS tags, following (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

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

the sequence of the given POS tags, we stick to (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, a special case of XP, to denote that we cannot determine the head of the phrase because of an error in word order. As an example, consider the sentence **I ate lunch was delicious.* (*The lunch I ate was delicious.*). According to the superficial forms and local contexts, the phrase *I ate lunch* forms the S:



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



2.2.5 Fragments

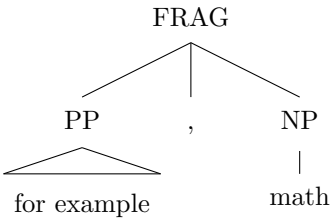
In learner corpora, sentences are sometimes incomplete in that the main verb of their main clause or the main clause itself is missing. They are called fragments. Examples are:

For example, math. (missing main verb)

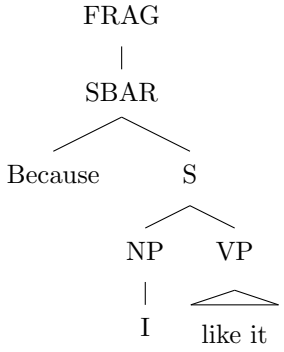
and

Because I like it. (missing main clause).

Fortunately, there already exists a tag for fragments in PTB-II: FRAG. Accordingly, we use it in our annotation scheme as well. For instance, the above examples give the parses:



and

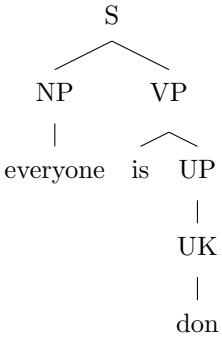


An exception is incomplete sentences which are defined as S in the bracketing guidelines for biomedical texts (Warner, Lanfranchi, O'Gorman, Howard, Gould, and Regan 2012). We tag such incomplete sentences as S following the convention. For example, an adjective phrase can form an S (e.g., (S (ADVP Beautiful)!)).

2.2.6 Unknown Words and Phrases

There are cases where one cannot determine the tag of a given word. We use the tag, UK, 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 introduced in Section 2.2.2. Based on this, the above example gives the parse:



For a phrase whose head word is unknown because of some error(s) in it, we use the XP tag instead of the UP tag. We use the XP-ORD tag to put them together into a phrase.

3 Corpus Annotation

We show the consistency of our new phrase structure annotation scheme using two different language learner corpora, and we investigate in detail characteristic CFG rules between native speakers and language learners of English.

3.1 Dataset

We used the Konan-JIEM (KJ) learner corpus (Nagata et al. 2011) as our primary dataset. The corpus contains essays written by college students whose English proficiency ranges from befinner to intermediate. The KJ corpus has annotations for POSs, chunking, and grammatical errors and their corrections. This information was shown to annotators when they annotated the phrase structures. To show the consistency of the annotation scheme, we also selected another language learner dataset called the International Corpus Network of Asian Learners of English (ICNALE) corpus (Ishikawa 2011). Although we used a subset of the ICNALE corpus, it has a larger variety of learners than the KJ corpus with respect to proficiency (beginning to advanced levels[5]) and native language (China, Indonesia, Japan, Korea, Taiwan, Thailand, Hong Kong, Singapore, Pakistan, and the Philippines). Table 2 shows some basic statistics on the two learner corpora. We note that the annotations are applied to original sentences as in Berzak et al. (2016). Because the KJ corpus contains learner errors and the respective corrections, it allows us to use them for the phrase structure annotation.

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

3.2 Annotation Procedure

Two professional annotators[6] participated in the annotation process. The first annotator helped revise the annotation scheme through discussions with the authors. Then, both annotators annotated a subset of the KJ corpus (11 texts, 955 tokens), where they were allowed to consult each other. For computing their inter-annotator agreement rate, another subset of the KJ corpus and a subset from the ICNALE corpus (59 texts, 12,052 tokens) was annotated by the annotators independently. The first annotators annotated the rest of the KJ corpus.

Table 3 shows the inter-annotator agreement measured in recall, precision, F-measure, complete match rate, and chance-corrected measure (Skjærholt 2014).[7] We used the EVALB tool[8] with the evaluation parameter proposed by Collins (1997), where we regard the annotation results of the first annotator as the gold standard set. We also used the syn-agreement tool[9] to calculate chance-corrected measure. The results show 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. A comparison between the results of the two annotators shows the major sources of disagreement. 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

Table 2 Statistics on annotated learner corpora. We used a subset of the ICNALE corpus

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 3 Inter-annotator agreement measured using recall (R), precision (P), F-measure (F), complete match rate (CMR), and chance-corrected measure (CCM)

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

[6] 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.

[7] As shown by Mizumoto, Hayashibe, Komachi, Nagata, and Matsumoto (2012), the KJ corpus has a higher error ratio than other learner corpora. Because the high error rate may deteriorate the consistency of annotations, we computed the inter-annotator agreement in order to see how the annotation guideline helps the consistency.

[8] http://nlp.cs.nyu.edu/evalb/

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

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.

3.3 CFG Rules in Learner English

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 determine if the annotation results represent the 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);[10] 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 $p(B|A)$ using the expected likelihood estimation, which is equivalent to Laplace smoothing with alpha = 0.5.[11] 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 pre-terminal symbols to focus on the structural characteristics. In addition, it excludes rules containing a Quantifier Phrase (QP; e.g., (NP (QP 100 million) dollars)), which frequently appear and form one of the characteristics of 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 $PP \rightarrow \phi$

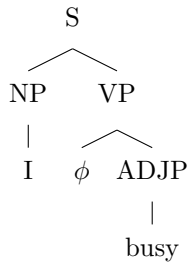
[10] To confirm that the extracted characteristics are not influenced by the differences in the domains of the two corpora, we also compared the learner data with the native speaker sub-corpus in ICNALE that is in the same domain. It turned out that the extracted CFG rules, were very similar to those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

[11] We provide an analysis to determine common characteristics in some learners groups (e.g., native languages and learner proficiency). This method is not constrained to analyzing rules regardless of error tags, and therefore it enables us to extract phrase structures that frequently appear in learner corpora. It is possible to focus on rules that are error-tagged (i.e., direct analysis), and this is one of the interesting extensions of this work.

Table 4 Characteristic CFG rules

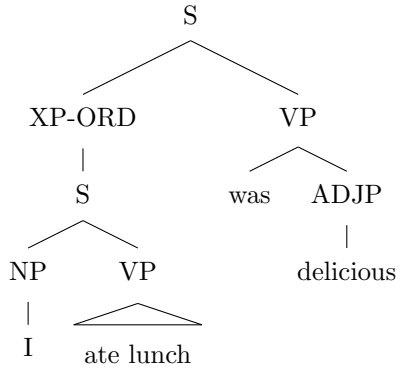
Overuse	Score	Underuse	Score
PP → ϕ NP	9.0	NP → NP , NP ,	−4.6
S → XP VP .	7.2	S → NP NP	−2.7
PP → IN IN S	6.7	S → NP VP . ”	−2.6
S → XP .	6.6	ADVP → NP RBR	−2.5
VP → ϕ ADJP	6.5	S → S , NP VP .	−2.4
VP → ϕ NP	6.3	NP → NP , SBAR	−2.4
SBAR → IN NN TO S	6.1	SBAR → WHPP S	−2.3
PP → ϕ S	6.1	VP → VBD SBAR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

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* → 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* → 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 appears in a sentence.

The second rule *S* → *XP VP .* implies that the subject NP cannot be recognized because of a combination of grammatical errors (c.f., *S* → *NP VP* .). The corpus data show that 21% of *XP* in *S* → *XP VP .* are actually *XP-ORD* concerning an error in a relative clause, as in:



(which was already shown in Section 2.2.4). Some of the learners apparently have problems with 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 example, the eighth rule in Table 4 $VP \rightarrow VBD\ SBAR$ 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: the Stanford Statistical Natural Language Parser (ver. 2.0.3) (de Marneffe, MacCartney, and Manning 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

Table 5 Parsing performance on learner English

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

corpora despite the fact that it contains a number of grammatical errors as well as 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)).

Figure 3 shows the relationship between sentence lengths and parsing performance using the F-measure. The horizontal axis corresponds to sentence length bins whose width is set to five; for example, the first bin (= 0) corresponds to sentences whose lengths are less than five. The solid and dashed lines represent the average of F-measure and its 95% confidence intervals. Note that in Figure 3, F-measure is defined as zero for sentences whose recall and precision are both zero.

As expected, Figure 3 clearly shows that longer sentences degrade parsing performance, which ranges from approximately 90% to 75%.[12] An exception is the first bin (sentence lengths of less than five) whose F-measure is approximately 80% in Figure 3; one would expect from Figure 3 that its performance should be more than 90%. Looking into the actual parsing results reveals that fragments consisting of one or two words such as "*Thirdly .*" are often mistakenly parsed as an interjection as in "(INTJ (ADVP (RB Thirdly)) (. .)))" (correctly: "(FRAG (ADVP (RB Thirdly)) (. .))"). These fragments degrade both recall and precision performance.

To achieve further improvement, we augmented the Charniak-Johnson parser with the learner data. We first retrained its parser model using sections 2–21 of the Penn Treebank Wall Street

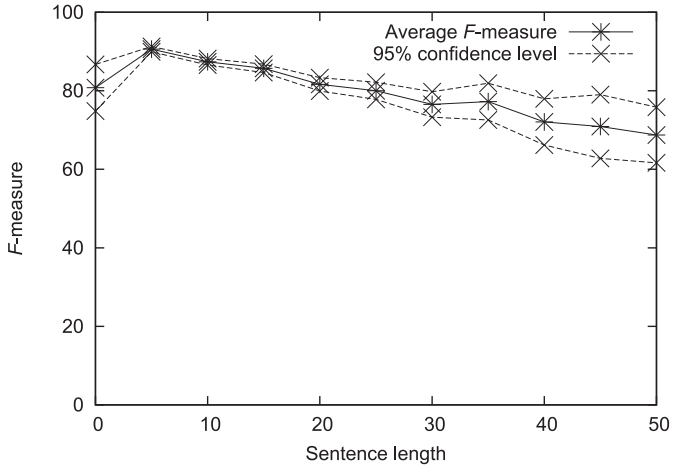


Fig. 3 Relation between sentence length and parsing performance

[12] Note that the 95% confidence intervals are wider for longer bins because instances falling into these bins tend to be fewer.

Journal (hereafter, WSJ) as training data and its section 24 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 the data into six parts, each of which consisted of approximately 61 essays, used one sixth as test data, another sixth as development data instead of section 24, 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 sections 2–21 of the 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.

Figure 4 shows the results. The simple addition of the learner data ($n = 1$) already outperforms the parser trained only on sections 2–21 of the WSJ ($n = 0$) with respect to both recall and precision, achieving an F-measure of 0.866 and a complete match rate of 0.515. The augmented parser model works particularly well at 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), and 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 weight n increases (up to $F = 0.878$ when $n = 24$), which shows the effectiveness of using learner corpus data as training data.

We can summarize the findings as follows; (i) the state-of-the-art phrase structure parsers for native English are effective even when parsing learner English; (ii) these parsers are successfully augmented by learner corpus data; (iii) the evaluation results support a previous report

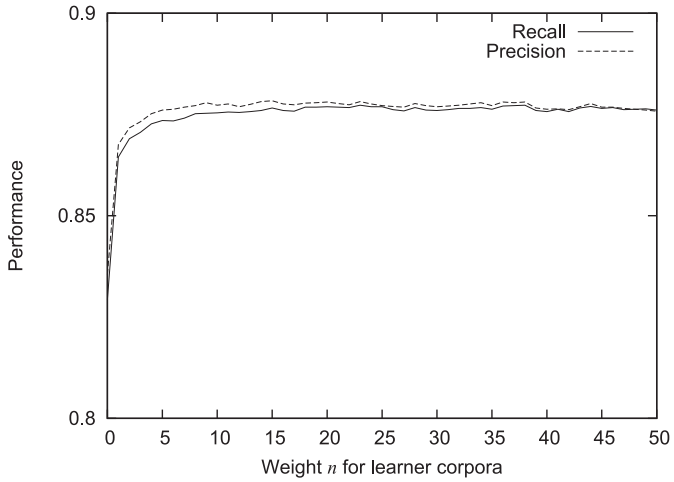


Fig. 4 Relation between learner corpus size in the training data and parsing performance

(Tetreault, Foster, and Chodorow 2010) that they are effective at extracting parse features for grammatical error correction (and probably for related NLP tasks such as automated essay scoring).

5 Conclusions

In this paper, we presented the first *phrase structure* annotation scheme for learner English. The scheme consists of five principles to deal with noisy learner English. We annotated two different learner corpora using it and showed that it was effective at consistently annotating learner corpora with phrase structures (i.e., it has a high inter-annotator agreement rate). We also investigated characteristic phrase structures in the learner corpora, and reported on phrase structure parsing performance on learner English for the first time. The annotation guidelines, annotated data, and parsing model for learner English created in this work are now available to the public.[13]

In our future work, we will evaluate parsing performance on other learner corpora such as ICLE (Granger, Dagneaux, Meunier, and Paquot 2009), which consists of a wide variety of learner English. We will also extend phrase structure annotation, especially working on function tags.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable feedback. A part of this paper is published at the 54th Annual Meeting of the Association for Computational Linguistics (Nagata and Sakaguchi 2016). This work was partly supported by Grant-in-Aid for Young Scientists (B) Grant Number JP26750091.

Reference

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). "Universal Dependencies for Learner English." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 737–746, Berlin, Germany. Association for Computational Linguistics.

[13] 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/.

Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. (1995). "Bracketing Guidelines for Treebank II-Style Penn Treebank Project.".

Charniak, E. and Johnson, M. (2005). "Coarse-to-fine N-best Parsing and MaxEnt Discriminative Reranking." In *Proceedings of 43rd Annual Meeting on Association for Computational Linguistics*, pp. 173–180.

Collins, M. (1997). "Three Generative, Lexicalised Models for Statistical Parsing." In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics*, pp. 16–23.

Dale, R. and Kilgarriff, A. (2011). "Helping Our Own: The HOO 2011 Pilot Shared Task." In *Proceedings of 13th European Workshop on Natural Language Generation*, pp. 242–249.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). "Generating Typed Dependency Parses from Phrase Structure Parses." In *Proceedings of 5th International Conference on Language Resources and Evaluation*, pp. 449–445.

Díaz-Negrillo, A., Meurers, D., Valera, S., and Wunsch, H. (2009). "Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT." *Language Forum*, **36** (1–2), pp. 139–154.

Dickinson, M. and Ragheb, M. (2009). "Dependency Annotation for Learner Corpora." In *Proceedings of 8th Workshop on Treebanks and Linguistic Theories*, pp. 59–70.

Foster, J. (2007a). "Treebanks Gone Bad: Generating a Treebank of Ungrammatical English." In *2007 Workshop on Analytics for Noisy Unstructured Data*, pp. 39–46.

Foster, J. (2007b). "Treebanks Gone Bad: Parser Evaluation and Retraining Using a Treebank of Ungrammatical Sentences." *International Journal on Document Analysis and Recognition*, **10** (3), pp. 129–145.

Foster, J. and Andersen, Ø. E. (2009). "GenERRate: Generating Errors for Use in Grammatical Error Detection." In *Proceedings of 4th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 82–90.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English v2.* Presses universitaires de Louvain, Louvain.

Ishikawa, S. (2011). *A New Horizon in Learner Corpus Studies: The Aim of the ICNALE Project*, pp. 3–11. University of Strathclyde Publishing, Glasgow.

Izumi, E., Saiga, T., Supnithi, T., Uchimoto, K., and Isahara, H. (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), pp. 119–125.

Kimura, M. (2003). "Japanese EFL Learners' Process of Noun Phrase Development: A Performance Analysis Using L2 Learners' Spoken Data." *Journal of Educational Research*, **8**,

pp. 61–67.

Mizumoto, T., Hayashibe, Y., Komachi, M., Nagata, M., and Matsumoto, Y. (2012). "The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings." In *Proceedings of COLING 2012: Posters*, pp. 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Nagata, R. and Sakaguchi, K. (2016). "Phrase Structure Annotation and Parsing for Learner English." In *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1837–1847.

Nagata, R., Whittaker, E., and Sheinman, V. (2011). "Creating a manually error-tagged and shallow-parsed learner corpus." In *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1210–1219.

Narita, M. (2013). "Analyzing Noun Phrase Postmodifiers in the Konan-JIEM Learner Corpus (in Japanese)." *Journal of Tokyo International University. The School of Language Communication*, **9**, pp. 1–12.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). "The CoNLL-2013 Shared Task on Grammatical Error Correction." In *Proceedings 17th Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12.

Petrov, S. (2010). "Products of Random Latent Variable Grammars." In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27.

Ragheb, M. and Dickinson, M. (2012). "Defining Syntax for Learner Language Annotation." In *Proceedings of 24th International Conference on Computational Linguistics*, pp. 965–974.

Ragheb, M. and Dickinson, M. (2013). "Inter-annotator Agreement for Dependency Annotation of Learner Language." In *Proceedings of 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 169–179.

Skjærholt, A. (2014). "A Chance-corrected Measure of Inter-annotator Agreement for Syntax." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 934–944.

Tetreault, J., Foster, J., and Chodorow, M. (2010). "Using Parse Features for Preposition Selection and Error Detection." In *Proceedings of 48nd Annual Meeting of the Association for Computational Linguistics Short Papers*, pp. 353–358.

Warner, C., Lanfranchi, A., O'Gorman, T., Howard, A., Gould, K., and Regan, M. (2012). "Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines.".

Keisuke Sakaguchi: Keisuke Sakaguchi is a Ph.D. student at Johns Hopkins University, Center for Language and Speech Processing. He received his M.E. degree from Nara Institute of Science and Technology in 2013, M.A. from University of Essex in 2006, and B.A. from Waseda University in 2005. He has been working on robust algorithms for non-canonical text processing.

Ryo Nagata: Ryo Nagata graduated from the Department of Electrical Engineering, Meiji University in 1999, completed the doctoral program in information engineering at Mie University in 2005 and became a research associate at Hyogo University of Teacher Education. Since 2008, he has been an associate professor at Konan University. His research interests are language modeling, grammatical error detection and correction, and edu-mining (educational data mining). He is a member of the Institute of Electronics, Information and Communication Engineers.

(Received November 16, 2016)
(Revised February 2, 2017)
(Accepted March 22, 2017)