Automatic Detection of Grammatical Structures from Non-Native Speech

Suma Bhat[1]*, Su-Youn Yoon and Diane Napolitano*[2]

[1]Beckman Institute, University of Illinois, Urbana-Champaign, USA
[2]Educational Testing Service, Princeton, USA
spbhat2@illinois.edu, {syoon,dnapolitano}@ets.org

Abstract

This study focuses on the identification of grammatical structures that could serve as indices of the grammatical ability of non-native speakers of English. We obtain parse trees of manually transcribed non-native spoken responses using a statistical constituency parser and evaluate its performance on noisy sentences. We then use the parse trees to identify the grammatical structures of the Index of Productive Syntax (IPSyn), previously found useful in evaluating grammatical development in the context of native language acquisition. Empirical results of this study show: a) parsing ungrammatical sentences using a probabilistic parser suffers some degradation but is still useful for further processing; and b) automatic detection of the majority of the grammatical structures measured by IPSyn can be performed on non-native adult spoken responses with recall values more than 90%. To the best of our knowledge, this is the first study which explores the relationship between parser performance and the automatic generation of grammatical structures in the context of second language acquisition.

Index Terms: syntactic parsing, non-native speech, grammatical development, grammatical error in speech, second language acquisition

1. Introduction

Automatic speech scoring of learner language involves assigning spoken responses a score of language ability taking into account the dimensions of fluency, intonation, pronunciation and grammar (e.g. [1]). In addition to scoring for language ability, providing feedback on what the learner is expected to know and what the learner does well, helps the learner improve his/her language performance. This study focuses on grammatical structures that could serve as criteria for feedback on grammatical ability and the degree to which they can be detected automatically.

Indices of grammatical ability, owing to their role as correlates of the developmental (and degenerative) process in humans, have played a critical part in the areas of child language acquisition (CLA), and language/cognitive impairment [2, 3]. In the domain of CLA they serve as indicators of specific milestones in formative grammar development; e.g. Developmental Sentence Scoring (DSS) [4], the Developmental Level (D-Level) [5] and the Index of Productive Syntax (IPSyn) [6]. In each of these cases, the index is obtained by noting the occurrence, within a language sample, of a number of grammatical structures that correspond to the complexity of language.

Syntactic complexity refers to the range and degree of sophistication of grammatical forms that surface in language production and has been found to be useful in characterizing the grammatical ability of language learners [7, 8, 9, 10]. It is conceivable, that the indices of grammatical ability mentioned above could serve as measures of syntactic complexity and be useful in the domain of second language acquisition (SLA) for the same reasons that they are useful in CLA. From a practical standpoint, knowing the trajectory of acquisition of grammatical structures in SLA, permits the creation of a list of those structures, which can then be used to generate learner feedback in computer-aided language learning or automatic scoring scenarios. We take the first step in this direction by choosing the grammatical structures of IPSyn (explained in Section 2.1) and studying the automatic identification of the associated grammatical structures from spoken responses. Given the exploratory nature of our study, we use manual transcriptions of spoken responses with disfluencies and repairs removed and use probabilistic parsers to detect the grammatical structures of interest.

Non-native language, owing to its inherent idiosyncratic constructions and grammatical errors, poses specific challenges to NLP tools, which includes syntactic parsers. This affects the downstream processing of the parse yield, which in our case is the automatic identification of grammatical structures of IPSyn. Consequently, a portion of our study is devoted to analyzing the extent of this degradation.

The purpose of this paper is two-fold. First, we investigate the extent to which a statistical parser designed primarily for native English-language writing can be used with non-native spoken responses. Second, we explore the extent to which the grammatical structures of IPSyn can be detected from parse trees of non-native spoken sentences. Empirical results of our study show the following.

1. The syntactic parses of ungrammatical sentences using a state-of-the-art probabilistic parser suffer some degradation but are still useful for downstream processing.
2. Automatic detection of a majority of the grammatical structures that are part of IPSyn can be reliably performed on non-native adult spoken responses with recall values more than 90%.

With its focus on the detection of *correct language-specific* production in non-native responses, this study sets itself apart from related studies that focus on the detection of erroneous production (e.g. [11, 12, 13]). To the best of our knowledge, this is the first such study which explores the relationship between parser performance and automatic IPSyn feature generation on non-native spoken responses.

2. Related Prior Work

In the field of SLA, syntactic complexity measures have been found to be useful for quantifying differences in grammatical ability across different proficiency levels, both for the purpose of language assessment and to capture the rate of longitudinal grammar development in second language writing [8, 9, 10].

Recently, with the advent of automated systems of language assessment, measures of syntactic complexity have been studied for spoken and written responses [14, 15]. The measures have been reliant on either part-of-speech- or clause/sentence length-based information but not on specific grammatical structures.

Several studies in the area of CLA have used sets of grammatical constructions as indicators of specific milestones in grammar development, including [5, 6]. The D-Level scale [5, 16] classifies sentences into levels according to the presence of particular grammatical expressions and mainly identifies simple and complex sentences, such as those which use subordinate clauses, subordinating conjunctions, or non-finite clauses in adjunct positions.

There have also been a few studies which have investigated the relationship between ungrammatical sentences and wide-coverage probabilistic parsers. In [17], the focus was on the evaluation of the Charniak parser's ability to produce an accurate parse for an ungrammatical sentence. Using a corpus of ungrammatical sentences and their corrected forms from written-language sources, the Charniak parser demonstrated solid performance via perfect F-scores on nearly a third of the sentences with grammatical errors. Additionally, they found that agreement errors and the use of the wrong preposition did not significantly affect the parser's output as compared to other kinds of errors.

The first part of our study is most similar to [17] in terms of the goal and the method of evaluation. However, the fact that our analysis accounts for a wider set of grammatical errors and is based on a corpus of non-native spoken responses, with their idiosyncratic constructions and particular types of grammatical errors, sets it apart from previous studies.

2.1. The Index of Productive Syntax

Similar to the D-Level scale, the Index of Productive Syntax (IPSyn), captures improvements in the grammatical ability of native English-speaking children during the early stages of language acquisition [6]. Towards this, IPSyn provides a well-organized inventory of grammatical forms: 60 structures consisting of 12 which are noun-based, 17 which are verb-based, 20 which are sentence-based, and 11 which examine questions and the use of negations[1]. The structures vary from simple constructs, such as noun phrases, to more complex constructs, such as bi-transitive predicates, conjoined sentences, and infinitive clause construction (the structures are numbered in increasing order of complexity). For each response spanning several sentences, a score of 0 (non-occurrence), 1 (occurring once) or 2 (occurring at least twice) is assigned to every structure of the inventory. The total score is then the sum of the scores over the 60 structures. Thus, the IPSyn score for a response is indicative of the diversity of grammatical expressions within it. In addition, noting that IPSyn was designed "as a measure of the emergence of syntactic and morphological abilities but not their mastery (p. 22)" [6], we would like to point out that the IPSyn score does not account for the accuracy of the grammatical expressions, it only accounts for the use of the structures.

Since IPSyn identifies syntactic complexity in terms of the diversity of production-based syntactic structures and not based on the length of individual sentences or clauses, the index seems best suited for SLA studies concerned about the range and sophistication of grammatical structures. An additional advantage comes with the fact that IPSyn was designed to be a flexible measure that captures the proper use of language, focusing on distinct forms of grammatical types and not tokens (as in, the frequency of occurrence of the forms). This, combined with the apparent scarcity of studies utilizing it in English SLA implies high, yet unexplored, potential of this measure of syntactic complexity in this domain.

From the point of view of SLA, an important question is whether IPSyn, designed to measure syntactic productivity in young children, is likely to be a useful index of L2 ability in adults. Despite its usefulness in early child language development, it is not usable for older children with normal language development who are well past the age of acquiring the basic language structures. However, a quick look at the IPSyn scores for a sample from our dataset of non-native English learners showed that the case was different for second language. Based on an initial analysis of the responses of 20 speakers (a subset from the highest and the lowest proficiency group), we found that not all speakers used S10 (adverbial conjunction) and S11 (prepositional complement), for instance. In particular, only a few speakers in the lowest proficiency group used them, whereas most speakers in the highest proficiency group used them. This suggests that, SLA, unlike early child language development, perhaps follows a different trajectory and that perhaps, IPSyn scores may be able to discriminate between stages of grammar development in our target population. Thus, from a theoretical standpoint, grammatical structures considered in IPSyn would be a natural starting point to investigate grammar development stages in SLA and automated methods such as ours would enable such an investigation.

Automatic systems with the use of such NLP tools as automatic part-of-speech taggers and syntactic parsers, designed to generate IPSyn scores, have been studied in [18, 19]. Both systems achieved a reasonably high accuracy in the identification of grammatical structures from native English-speaking children's spoken data; 93% with [18] and 97% with [19] (despite a 10% reduction in F-score owing to parsing errors). Their empirical results show that automatically-generated IPSyn scores are robust to errors made by the parser over child language samples, with [18] using its own dependency parser and [19] using the Charniak parser.

In [20], automatically derived IPSyn-based scores of syntactic complexity were found to be good predictors of proficiency scores for non-native children suggesting that these basic syntactic structures are discriminative between proficiency groups in children. However, we need more experiments to see if this is the case in a broader SLA context as well. Moreover, the accuracy of such an automated system for use on non-native spoken responses is not yet known.

In this study, we will focus on the impact of grammatical errors on both the parser and the automated detection of IPSyn structures. We speculate that some grammatical errors have a significant impact on the output from automatic parsers and would therefore cause problems in the detection of IPSyn structures, whereas other grammatical errors may affect the detection to a much smaller extent.

3. Data

Our original dataset contains 444 non-native spoken responses from 360 test takers of an international test of English, elicited as spontaneous responses to a set of questions that the subjects either read or listened to. All responses were scored for proficiency (reflecting delivery, language use, and content) by

[1]For a detailed description of the grammatical structures we direct the reader to the associated reference, but provide a brief description of the relevant structures in Section 5.2

human raters on a scale of 1-4, with higher scores indicating better speaking proficiency. The responses were manually transcribed verbatim, then later annotated for grammatical errors and their corrections, within the intended context. Disfluencies (e.g., fillers, false starts, repetitions, and self-repairs) were removed during this annotation process. Only one corrected form was suggested by the annotators per ungrammatical form.

Thus, every sentence of the dataset had three "forms" - the *original form* with disfluencies and grammatical errors, the *no disfluency form* where disfluencies were removed but grammatical errors (when present) were retained, and the *corrected form* without disfluencies and with the grammatical errors corrected. For the purpose of this study, we use the *corrected form* sentences as our gold standard, and only compared the *no disfluency* forms to their *corrected* forms.

	1	**2**	**3**	**4**	**TOTAL**
Entire set	21	335	760	286	1402
GrammarError set	16	231	492	107	846
OneError set	7	90	213	61	371

Table 1: Description of datasets used in this study. The totals indicate the number of sentences in each set by the proficiency scores assigned to them (1-4).

This pre-processing yielded a subset of the data (the *entire* set) with 325 speakers and their 395 responses, giving us a total of 1402 sentences with an average sentence length of 15 words. From this, we created a subset which includes the 846 sentences that have at least one grammatical error (the *GrammarError* set), of which 371 sentences had only one grammatical error (the *OneError* set). The number of sentences in each set is shown in Table 1 along with the distribution of proficiency scores within each of them. Since we only rated proficiency scores at the response-level, we used the proficiency score of the response which the sentence belongs to as proficiency score of the sentence. We notice that the data set is skewed in favor of responses from the higher proficiency groups, but we do not make a proficiency distinction in processing the sentences.

The distribution of the number of errors in the *GrammarError* set was skewed to the right, with mean and median values of two errors per sentence, and a mode of one error per sentence. The grammatical sentences in the *Entire* set had a mean length of 14.8 words and a standard deviation of 8.12 words.

4. Tasks

4.1. Parser evaluation

The Stanford Parser's English PCFG model [21] was trained on the Wall Street Journal and was used to parse the sentences in the *GrammarError* set. This parser was chosen for two reasons: for its ability to produce a parse for all sentences, regardless of their grammaticality; and for it being a competitive probabilistic phrase-structure parser. We assume that the *corrected* and *no disfluency* forms without grammatical errors are similar to formal English once disfluencies have been removed from the latter.

For the purpose of this study, we assume that the parses of the *corrected* (gold) sentences are the gold parses, and we compare these with those of the corresponding ungrammatical sentences. Since parser outputs on sentences containing grammatical errors that result in insertions, substitutions, and deletions are not guaranteed to be the same as those of their corrected versions, we analyze the differences in parse trees for each (*corrected*, *nodisfluency*) pair using the Sparseval labeled precision/recall measures [22]. Sparseval was designed to evaluate the mismatched parse yields of word hypotheses from an automated speech recognition system. For each pair, we use NIST's SCLITE [23] to establish alignments between comparable constituent spans for labeled bracketing scoring. We use this comparison-based evaluation to obtain precision, recall, and F-measure scores for the labeled bracketing.

4.2. Detection of IPSyn Structures

We use a rule-based system similar to the AC-IPSyn system [19] to identify the IPSyn structures using the output of the Stanford Parser. The system first identifies the IPSyn syntactic structures based on corresponding patterns matching POS tags and/or constituencies in the generated parse trees. Of the 60 structures included in IPSyn, we include all but the 11 question and negation structures, owing to the occurrence of only declarative sentences in our responses. For each IPSyn structure considered, we develop regular expressions intended to catch the construct being measured using part-of-speech and parse information, then make a binary decision: 0 if the regular expression did not match anywhere in the parser output, or 1 if it matched at least once. A point to note here is that, unlike the original response-level score of IPSyn, a sentence-level binary scoring is adopted in this paper since we evaluate the automatic detection of IPSyn structures and not the accuracy of IPSyn score prediction. From this, we create a set of 49 binary features per sentence.

We could not evaluate our IPSyn system for its accuracy in detecting grammatical structures from grammatical sentences due to the fact that we lack a corpus of sentences with manually-assigned IPSyn scores. Instead, we refer to the results in [19] as described in Section 2.1 that observed an accuracy of 97% in identifying IPSyn grammatical structures.

5. Results

5.1. Effect of grammatical errors on parser performance

We use the *GrammarError* set to evaluate the effect of grammatical errors on the quality of the generated parse tree. We calculate labelled bracketing precision and recall scores on the parse trees of the ungrammatical sentences. We also report the proportion of ungrammatical sentences whose parse tree matched the gold parse as a measure of the parser's efficacy (**Match**). The impact of grammatical errors is further reflected in the proportion of *GrammarError* sentences whose recall scores were lower than 75% (**Problematic**). These results are summarized in Table 2.

Despite the presence of grammatical errors, the parser achieved an overall recall score of 85% and a precision score of 84%, with 100% F-score on 42% of the ungrammatical sentences. This suggests that the parser performance is reasonable despite the multiple grammatical errors present in many of the sentences.

With exactly one grammatical error per sentence, we observed that in almost 60% of the cases the parse of the ungrammatical sentence and its grammatical counterpart matched whereas in about 16% of the sentences the recall was less than 75%. This suggests that different kinds errors affect the production of parse trees differently. Using a coarse classification of errors resulting from the choice of an incorrect word (for instance the wrong article or preposition), an extra word or a missing word, we observe that the impact of an incorrect word form (substitution) is far less compared to that of an extra word

	Precision	Recall	Match	Problematic
Overall	84.20	84.56	42.23	25.62
One Error	90.13	90.80	59.89	15.72
Two or more	79.59	79.71	28.48	33.33
Incorrect word	96.11	96.12	78.64	6.80
Extra word	84.25	88.60	39.29	23.21
Missing word	88.22	86.00	43.90	19.51
Level 1	85.71	80.96	50.00	31.25
Level 2	82.95	82.74	45.45	29.43
Level 3	84.59	85.11	40.12	24.24
Level 4	84.95	86.54	43.81	22.86

Table 2: The effect of grammatical errors on parser performance shown in terms of labeled bracketing precision, recall, percent matching gold and *GrammarError* pairs (**Match**), and percent gold and *GrammarError* pairs with recall less than 75% (**Problematic**). The numbers on the P and R columns are the mean values, but owing to their distributions being highly left-skewed, the other two columns are more informative. We include the performance of the case of coarsely classified sentences with one grammatical error (*incorrect, extra* or *missing* word) as well as that based on the proficiency score of the sentences (1,2,3 or 4).

(insertion) or a missing word (deletion). In general, substitution results in an F-score of 96%, and more than 75% matches, whereas insertion and deletion result in lower F-scores (86%) and matches, with higher proportions of parses with low recall. Not surprisingly, we see a drop in the recall/precision scores and the proportion of complete matches but an increase in the proportion of problematic cases as the number of errors increase.

R: There are two main **important** reasons why I am of this opinion.
WM: I have taken a music class before and I really enjoyed **it** a lot. (*it* was omitted)
V: The professor **explain** the suitability of animals for domestication.
VT: One group was told they **will** be watched.
PREP: So it's better to live **in** campus.
LE: The other definition is the broader one which means we'll use money to **do** purchases.
A: Take the example of **a** taxi driver. (The article *a* was omitted)
N: First year **student** need to live in the dormitory on campus.

Table 3: Examples of ungrammatical sentences from the corpus with their tagged errors. The errors are in **bold**.

Next, in order to understand the impact of grammatical errors on parser performance, we focus on specific types of errors found in sentences containing only one grammatical error. For this purpose, we limit our attention to only those errors that occur in at least 10% of the corpus. The errors considered are those most commonly present in non-native responses [24]: article (A), preposition (PREP), verb tense (VT), and noun form error (N), verb form error (V), and also other errors prevalent in our corpus: word missing (WM), wrong lexical choice (LE) and redundant word (R) errors. Table 3 has sample sentences for each error. The results are summarized in Table 4.

Based on these results, missing word errors (WM) cause the most degradation of parser performance and general noun errors (N) the least. The N error category is concerned with the use of countable/uncountable nouns and with the number/morphology of nouns. WM errors include the omission of phrasal components and important function words, such as pronouns and conjunctions, but exclude article and preposition omissions, since these are captured by the A and PREP categories, respectively.

	WM	R	PREP	A	LE	VT	V	N
Precision	77	68	90	94	94	93	94	96
Recall	71	78	92	93	93	94	94	96
Match	16	6	59	73	61	70	73	84
Problematic	50	44	14	11	16	7	12	7
Corpus count	95	72	150	413	142	172	103	292

Table 4: Labelled bracketing precision and recall (rounded to the nearest integer), percentage of complete matches, and percentage of problematic cases for sentences with one grammatical error ordered by recall values. We restrict our analysis to errors that occur at least 15 times in the *OneError* set. "Corpus count" indicates their overall occurrence in the corpus (including sentences with multiple errors).

Within our limited sample, we observe that WM errors render sentences incomplete and thus affect the generated parse to a large extent. The omission of syntactically-central material, such as the finite verb (captured by WM), affects phrases structures the most, while sentences with other errors–for instance, agreement –can still be parsed in a robust manner. Moreover, with recall scores being higher than precision for R, PREP, and VT suggests that these error types introduce structures not found in the gold parse. Agreement errors and verb tense errors resulted in over 70% of cases having a completely matching parse tree, suggesting that the effect of these errors is relatively minor. We noticed that, in a majority of these cases, the part-of-speech tags remain unaffected by the presence of these errors within the sentence. Broadly, these observations, based on spoken non-native responses, are in line with those made on written responses (not necessarily non-native) in [17, 25, 26].

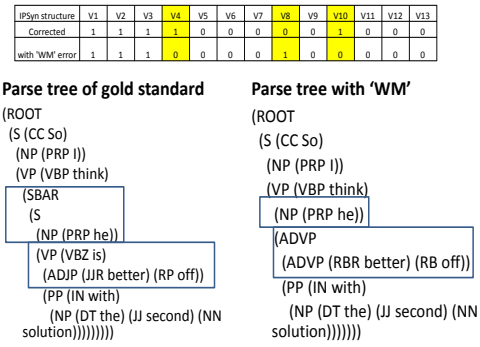

Figure 1: An example from our corpus to illustrate the effect of a WM error on the verb-related IPSyn structures (top) and labeled bracketing (bottom). Sentence: *So I think he* (**is**) *better off with the second solution.* The verb *is* is missing.

The effect of a WM error is illustrated in Figure 1. The missing verb form results in the loss of some constituents (SBAR and S) and introduces a new constituent (ADVP). This causes a misalignment of labeled brackets (shown in boxes), thus resulting in low recall.

5.2. Effect of parser performance on the grammatical structures used in IPSyn

Next we consider the extent to which the IPSyn grammatical structures can be detected from parses of ungrammatical sentences. We look at the precision and recall scores of detecting the structures, calculated with respect to the structures derived from the corresponding gold sentence, that occurred a minimum of 56 times (6%). This limits our analysis to 11 out of the 12 noun structures, 13 of the 17 verb structures, and 11 of the 20 sentence structures; a total of 35 out of our original subset of 49 structures considered. The remaining 14 structures either did not occur, or were not detected in the gold sentences. It is likely that the rules for detecting these structures using regular expressions over POS tags and constituents were sufficient for

	N5	N6	N7	N8	N9
P	90.56	92.45	95.24	92.31	87.80
R	86.78	89.77	81.08	82.35	88.32

Table 5: Precision and recall (in %) for noun-related structures in sentences with at least one grammatical error. Only those structures with recall lower than 90% are shown here. We restrict our analysis to the 11 of the 12 noun structures that occur at least 56 times in the corpus. **N5**: article before a noun, **N6, N8**: two-word noun phrase (article + noun) before and after a verb respectively, **N7**: plural noun form, **N9**: noun phrase of the form determiner + modifier + noun.

CLA but are inadequate for adult non-native speakers.

We calculate precision and recall by counting the true positives, false positives, and false negatives obtained by comparing the occurrence of structures in the parse trees of the sentences in their *no disfluency* form and their corresponding *corrected* form from the *GrammarError* set. The results are tabulated in Table 5 for the noun-related structures, Table 6 for the verb-related structures, and Table 7 for the sentence-related structures. Owing to space constraints only those structures with recall scores lower than 90% are tabulated. We note that those structures with recall values higher than 90% also had precision values greater than 90%.

From Table 5 we observe that the detection of 5 of the 11 noun structures is negatively affected by the presence of grammatical errors, with recall values lower than 90%. The recall of **N7** is the most affected and this can be directly attributed to general noun errors made by English-language learners. N5, N6, N8, and N9 are additionally impacted by the omission of articles, the most common error type (A) in this corpus (see Table 4). Despite their lower recall, these structures appear to be reasonably reliably detected as evidenced by their high precision scores. Additionally, the other 6 structures have a recall of at least 90%, indicating their robustness.

	V6	V7	V10	V11	V12	V16
P	82.24	82.19	85.93	95.18	90.63	90.91
R	83.33	81.63	88.86	71.82	76.32	66.67

Table 6: Precision and recall scores (in %) for the verb-related structures most impacted by the presence of at least one grammatical error. We restrict our analysis to 13 of the 17 verb structures that occur at least 56 times in the corpus. **V6**: auxiliary verb, **V7**: progressive suffix, **V10**: third person present tense suffix, **V11**: past tense modal verb, **V12**: regular past tense **V16**: past tense copula.

Similarly, for the verb structures, the recall of almost half of the structures (6 out of 13) are affected by the presence of grammatical errors. The precision and recall for these structures can be found in Table 6, where one may observe that the recall value of **V16**, **V12**, and **V11** suffered the most. The recall values of these structures can be directly attributed to the verb tense errors (VT) in the ungrammatical sentences. In addition, we notice that the recall of V6 , V7 and V10 are affected by the verb errors of the learners. Regardless, detection of the remaining 7 structures is robust, with both precision and recall greater than 94% (and hence not shown).

In the case of sentence-related structures, 7 out of 11 of the structures are robust with recall values greater than 90% and corresponding precision values greater than 90% (refer Table 7).

We summarize our observations on the robustness of the IPSyn structures in Table 8. From this table we notice that a majority of the structures (19 out of 35) are reasonably reliably-detected despite the presence of grammatical errors. With a recall threshold of 80%, however, all but 3 structures are reliably detected (these are not shown in the table).

	S10	S17	S14	S18
P	94.44	68.15	84.11	82.73
R	89.75	83.33	81.82	80.99

Table 7: Precision and recall scores (in %) for the sentence-related structures most affected by the presence of at least one grammatical error. We restrict our analysis to those structures that occur at least 56 times (11 of the 20 structures). **S10**: adverbial conjunction, **S14**: bitransitive predicate, **S17**: infinitive clause, **S18**: gerund.

	N-based	V-based	S-based	TOTAL
Recall <90%	6	6	4	16
Recall >90%	5	7	7	19
TOTAL	11	13	11	35

Table 8: Summary of robust (Recall >90%) and non-robust IPSyn structures (Recall <90%).

We then examine the cause of low recall for these structures, with the assumption that the degree to which a grammatical error impacts an IPSyn structure is reflected in its recall. Limiting our analysis to sentences with only one grammatical error allows us to isolate the effect of a particular error type. Here we focus on the error categories WM and R since we have observed that they have the most impact on parser performance (see Table 4), and separately on A, N, and PREP, since those account for most of of the errors in non-native language.

We find that the WM errors impact the recall values of the detection of 14 out of 35 structures, resulting in a recall of 50% for **S18**, **V11**, and **V7** (gerund, past tense, and progressive suffix, respectively). Consider the example provided in Figure 1, where the missing verb affects the detection of the verb-related structures **V4**, **V8**, and **V10**. On the other hand, A, N and PREP errors, which are, again, more prevalent in non-native responses than WM errors, do not affect many structures; but when they do, their effect is significant, resulting in recall values of 65.67%, 41.5%, and 50% in **N5**, **N7**, and **S14**, respectively.

6. Interpretation of Results

Empirical results of our experiments above suggest that the overall degradation in parser output is relatively low despite the prevalence of grammatical errors in the sentences, specifically a reduction of 15% in recall on the ungrammatical sentences. Despite this, we saw that a majority of the grammatical structures considered in the analysis, 32 out of 35, had recall values greater than 80%, with 19 of them having recall values greater than 90%. For automatic detection of IPSyn structures to potentially serve as a significant labor-saving option, a very high recall with reasonable precision is required. Our results show that this is true for 19 of the 32 structures and that these structures are also identified with high precision (greater than 90%) suggesting their potential of being used in response-specific feedback generation on language forms correctly produced. It remains to be explored whether the parser performance on ungrammatical sentences is better with a parser such as the Berkeley parser [27] which imposes no implicit linguistic constraints.

In this study, we used sentences corrected by native English annotators as the gold standard forms of their ungrammatical counterparts. For some sentences, however, it is likely that there are multiple ways of correcting the grammatical errors, especially when the intended meaning of the non-native speaker was not clear from the context. Such ambiguity has

been an important issue in a wide variety of NLP tasks relating to second language learners (e.g., error annotated learners' corpus construction, part-of-speech tagging, and parsing). Here, the minimum number of corrections needed to retain as many portions of the learner's original sentence was applied by the annotators. It is likely that the parser would be sensitive to the subjectivity in the grammatical error correction process, as well as on which grammatically-correct version was chosen (when more than one possible grammatical sentence exists).

Since a response consists of multiple sentences, a structure affected by grammatical errors in one sentence may still occur in other well-formed sentences, and thus be detected. We will investigate the effect of grammatical errors on IPSyn structures at the response level in a future study.

7. Conclusions and Future Work

We studied the performance of the Stanford probabilistic constituency parser for use with manually-transcribed non-native English spontaneous spoken responses. The parse trees of sentences with grammatical errors were compared with the parses of their manually-corrected forms. The presence of grammatical errors resulted in a labeled bracketing F-score reduction of 16%; however, about 40% of the ungrammatical sentences had an F-score of 100%. Looking at the effect of only one grammatical error in a sentence, we find that missing words (specifically nouns, pronouns, verbs, and adjectives) affect the parse yield more than article and preposition errors.

The usability of the parse yield is confirmed in the second part of our study where we explore the robustness of a set of grammatical structures which have been found to be reliable indices of syntactic complexity. The results are promising, showing that a majority of these structures can be detected with recall values more than 90% from the parse trees.

In the future, we would like to examine parsing behavior on non-native spoken responses with the Stanford parser trained on the Switchboard corpus and have manually-generated gold parses available for a more thorough comparison of parse yields. A natural extension to this experiment is to explore how an automatic disfluency detection system and a parser could be combined. For practical usage, we will also explore the efficacy of the proposed method with automatically-recognized responses.

8. References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[2] T. Solorio, "Survey on emerging research on the use of natural language processing in clinical language assessment of children," *Language and Linguistics Compass*, vol. 7, no. 12, pp. 633–646, 2013.

[3] H. Cheung and S. Kemper, "Competing complexity metrics and adults' production of complex sentences," *Applied Psycholinguistics*, vol. 13, no. 01, pp. 53–76, 1992.

[4] L. L. Lee, *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians.* Northwestern University Press, 1974.

[5] S. Rosenberg and L. Abbeduto, "Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults," *Applied Psycholinguistics*, vol. 8, pp. 19–32, 1987.

[6] H. S. Scarborough, "Index of productive syntax," *Applied Psycholinguistics*, vol. 11, no. 1, pp. 1–22, 1990.

[7] L. Ortega, "Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing," *Applied linguistics*, vol. 24, no. 4, pp. 492–518, 2003.

[8] C. P. Casanave, "Language development in students' journals," *Journal of Second Language Writing*, vol. 3, no. 3, pp. 179–201, 1994.

[9] S. Ishikawa, "Objective measurement of low-proficiency efl narrative writing," *Journal of Second Language Writing*, vol. 4, no. 1, pp. 51–69, 1995.

[10] K. Henry, "Early l2 writing development: A study of autobiographical essays by university-level students of russian," *The Modern Language Journal*, vol. 80, no. 3, pp. 309–326, 1996.

[11] S.-M. J. Wong and M. Dras, "Parser features for sentence grammaticality classification," in *Proceedings of the Australasian Language Technology Association Workshop*, 2010, pp. 67–75.

[12] J. Tetreault, J. Foster, and M. Chodorow, "Using parse features for preposition selection and error detection," in *Proceedings of the ACL 2010 Conference Short Papers*, 2010, pp. 353–358.

[13] M. Heilman, A. Cahill, N. Madnani, M. L. M. Mulholland, and J. Tetreault, "Predicting grammaticality on an ordinal scale," in *Proceedings of the ACL 2014 Conference Short Papers*, 2014.

[14] S. Bhat, H. Xie, and S.-Y. Yoon, "Shallow analysis based assessment of syntactic complexity for automated speech scoring," in *Proceedings of the ACL 2014 conference Long Papers*, 2014.

[15] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[16] M. A. Covington, C. He, C. Brown, L. Naci, and J. Brown, "How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale," CASPR Research Report 2006-01, Athens, GA: The University of Georgia, Artificial Intelligence Center, Tech. Rep., 2006.

[17] J. Foster, "Parsing ungrammatical input: an evaluation procedure," in *LREC*, 2004.

[18] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proceedings of the ACL 2005 conference*, 2005, pp. 197–204.

[19] K.-n. Hassanali, Y. Liu, A. Iglesias, T. Solorio, and C. Dollaghan, "Automatic generation of the index of productive syntax for child language transcripts," *Behavior research methods*, vol. 46, no. 1, pp. 254–262, 2014.

[20] K.-n. Hassanali, "Using natural language processing for child language analysis," 2013.

[21] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the ACL 2003 conference*, 2003, pp. 423–430.

[22] B. Roark, M. Harper, E. Charniak, B. Dorr, M. Johnson, J. G. Kahn, Y. Liu, M. Ostendorf, J. Hale, A. Krasnyanskaya *et al.*, "Sparseval: Evaluation metrics for parsing speech," in *Proceedings of the LREC conference*, 2006.

[23] J. G. Fiscus, J. Ajot, N. Radde, and C. Laprun, "Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech," in *The International Conference on language Resources and Evaluation (LERC)*, 2006.

[24] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The conll-2013 shared task on grammatical error correction," in *Proceedings of CoNLL*, 2013.

[25] J. Foster, J. Wagner, and J. Van Genabith, "Adapting a wsj-trained parser to grammatically noisy text," in *Proceedings of the ACL 2008 conference*. Association for Computational Linguistics, 2008.

[26] J. Wagner and J. Foster, "The effect of correcting grammatical errors on parse probabilities," in *Proceedings of the 11th International Conference on Parsing Technologies*. Association for Computational Linguistics, 2009.

[27] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing." in *HLT-NAACL*. Citeseer, 2007, pp. 404–411.