Tagging SVA Errors in L2 Writing: Implementation and Evaluation

Developed and written by James Taylor, Ella Alhudithi, Tom Elliott, and Sondoss Elnegahy

1. Introduction

It is not unusual to expect grammatical errors as a feature in Second Language (L2) writing. Among these, errors attributed to agreements between subjects and verbs (SVA) have largely been documented in applied linguistics research studies (Stapa & Izahar, 2010; Singh et al., 2017; Wee et al., 2010; Zheng & Park, 2013). What these studies have revealed is that the quantity and type of SVA errors is influenced by many factors, including learners' language proficiency. With this being the case, the motivation of this project is to develop a program that tags SVA errors written by L2 learners. Developing such a tagger would be instrumental in investigating the frequency of SVA errors, in both lower and higher proficiency levels of L2 learners. In other words, results obtained from the tagger will help to determine if such an error is an identifiable feature of one or both proficiency levels. The following sections illustrate the step-by-step procedure that was performed, beginning with defining the construct, followed by describing the corpus, designing the tagger, evaluating the tagger's performance, and lastly, ending by discussing the findings and their relevance to applied linguistics research.

2. Methods

2.1 Detecting SVA Errors. As documented in the literature (Biber et al., 1999; Ellis & Barkhuizen, 2005; Gass et al., 2013), a subject-verb agreement (SVA) error occurs when there is a mismatch between the number of subjects and verbs. That is, labeling an error as a SVA indicates that a singular subject is used with a plural verb or a plural subject used with a singular verb. Dulay et al. (1982) developed a taxonomy to categorize such errors into four types: omission, addition, misinformation, and misordering. Examples of these error types are: (1) '*she work*' [omitting the singular marker from the verb], (2) '*they thinks*' [adding the singular marker to the verb], (3) '*the books is*' [misinforming the quantity of the subject], and (4) '*she not has worked*' [misordering the negator of the verb]. Given the scope of the present project, the focus

was placed into SVA errors attributed to three categories, namely omission, addition, and misinformation. Table 1 demonstrates all errors targeted in this project. However, it is important to note that instances in which errors are not attributed to SVA were excluded. Examples of these are misspellings (e.g., *she thiks*) and misordering (e.g., *he has living been*). Similarly, errors relating to modal verbs (e.g., can and could) were also ignored since no number agreement is required.

Table 1

A list of SVA errors addressed in this project

Syntactic position	Word category	Description	Example	Error type
subject	noun	singular and plural (e.g., family and jobs)	many jobs has been created.	misinformation
	demonstrative pronoun	this, that, these, those	this have been replaced.	misinformation
	personal pronoun	i, she, he, it, you, they, we	she create many items.	omission
	possessive pronoun	mine, hers, his, yours, theirs, ours	hers were the best story.	misinformation
	relative pronoun	which, that, who, whoever	I saw a woman who work at the UN.	omission
	existential (there)	there	there were a party last week.	misinformation
	numerical	numbers	at least 50 has joined the team.	misinformation
verb	past tense	regular (e.g., worked and learned) irregular (e.g., was, were, had, did, ate and wrote)	we was students last year.	misinformation
	present tense	singular (e.g., am, is, has, does, works and learns) plural (e.g., are, have, do, work and learn)	the authors supports this idea.	addition

2.2 Corpus. To provide our program with authentic examples of learner writing, the present project used a sample taken from the ISU EPT Corpus (Iowa State English Placement Test). The texts in this corpus came from a university-wide test for international students

consisting of two timed writing prompts, a summary and an argumentative essay. In order to best diagnose the program's ability to find errors in subject-verb agreement, only the lowest-scoring texts were sampled. In addition, the untagged version of the corpus was selected for tagging, working with the Penn Treebank's tagset within CoreNLP. All sample texts were spell checked to allow easier focus on grammatical and other issues rather than misspellings.

2.3 Program. The developed SVA *Tagger* (2020)[1] was created to perform the following actions: (1) assigning tags to L2 writing samples, (2) identifying misuses of SVA, and (3) reporting numbers of SVA misuses. To check for these errors, it utilizes Stanford's coreNLP pipeline (Qi et al., 2020), followed by some rule based error checking. The tagger is written in Python 3.8 and utilizes the stanza library, an implementation of Stanford's coreNLP for Python. After reading the corpus text file and putting it through the pipeline (tokenize → pos → lemma → depparse), it iterates through each sentence. First it creates a 'forward' dependency list for the sentence from the information given from the depparse. This list will give all dependents for each governor instead of what depparse provides (the governor for each dependent). Then, it goes through the sentence again and checks for combinations of pos tags and forward dependencies to find some specific relationship. This relationship is then checked against an agreement matrix of verb POS tags and nominal POS tags (or in the cases of pronouns, specific words). A more technical explanation of this is available in the README of our *repository*[2]. It is then added to the proper correct or incorrect list. After doing this for each sentence, the tagger evaluates how many subject verb relationships were found, as well as how many were correct or incorrect.

3. Results

Several steps were performed to evaluate the accuracy of the developed tagger in deducting SVA errors written by learners of English. This evaluation process was completed in two phases: a manual tagging of SVA errors (i.e., gold-standard), followed by an accuracy analysis of the tagger performance (i.e., precision, recall, and f1-score). For both phases, the

same sample data was used, consisting of a total of 2,271 words collected from 10 essays written by L2 students. The first phase was carried out by three coders who manually and separately tagged SVA errors, using two values: '1' for target errors and '0' for none. This manual tagging has resulted in identifying a total of 11 SVA errors, in which nine of these were tagged by all coders. Following the manual tagging, a discussion on the two instances in which coders disagreed about was initiated to determine whether they were attributed to the target error type. Lastly, the first phase ended with performing the Krippendorff's alpha test due to its effectiveness in measuring agreements among many coders annotating a small sample size (Allen, 2017; Hayes & Krippendorff, 2007; Swert, 2012). The results of this statistical test revealed that the inter-rater agreement was $\alpha = .99$, indicating a higher rate of agreement.

After these errors were tagged manually, further steps were carried out to measure the accuracy of the developed tagger. This second phase involved using the tagger to deduct SVA errors written in the same sample data tagged by the coders. After running the data, numerical information was recorded concerning 'all SVA errors that were tagged', 'all SVA errors that were tagged correctly', all SVA errors that were tagged incorrectly', and 'all SVA errors that were not tagged'. Once these numbers were obtained, three accuracy measurements were calculated: precision, recall, and f1-score. As illustrated in Table 2, the findings revealed that the Tagger was moderately accurate in marking SVA errors found in the sampled corpus texts, reaching (.63) for the F1 score. Given the results accounted for precision (0.50) and recall (0.84), these indicated that the Tagger was highly accurate in not tagging other errors as SVAs.

Table 2

Accuracy rates of the develop Tagger

Information	Description	Result
all SVA errors tagged		28
SVA errors tagged correctly	true positives	11
SVA errors tagged incorrectly	false positives	11

SVA errors were not tagged	false negatives	2
precision	true positives divided by (true positives and false positives)	0.50 (50%)
recall	true positives divided by (true positives and false negatives)	0.84 (84%)
f1 score	2*((Recall * Precision)/(Recall + Precision))	0.63 (63%)

4. Discussion

As shown in the results, the performance of the developed Tagger in marking SVA errors was satisfactory. Given this accuracy rate, creating such a tagging program would be indeed helpful in research that targets L2 grammatical errors and uses machine assessment. Given the existing human rater bias and being prone to human errors, using such tagging tools could help achieve more accuracy in detecting errors as well as reducing human labor. In fact, although this project used only a small set of texts and three raters, the program detected two subject-verb agreement errors that the human raters missed. Due to these affordances, many opportunities would be provided for researchers to explore a wide range of L2 writing issues. In particular, this program, in combination with the CoreNLP tagging and dependency trees, works fairly well with lower-level learner writing containing many errors, both those being targeted and others that can cause difficulty for automatic detection.

While the program offers some promising and useful affordances, it contains some limitations in its present state of development. Specifically, it returned a large number of false positives for a variety of reasons. This was in part due to the limitations of the CoreNLP system, with some words or phrases (*iPads, All in all*) or dependencies being mis-tagged, such as in the following sentences.

Ex1. i (PRP) <--nsubj-- Pads (VBZ) in "*iPads as a part of a year-long study of e-readers and test group had high satisfaction with iPads*."

Ex2. All (DT) <--nsubj-- are (VBP) in "*All in all, iPads are good for student use in classrooms.*"

Ex3. it (PRP) <--nsubj-- s (VBZ) in "*Thus, social media is a very good communication to interact with people but on the other side see* ***its cons*** *also."*

However, many of the false positives were due to limitations in our program itself, particularly in its conceptualization rather than implementation. For the scope of this class project, we did not account for some of the more complex grammatical constructions that were found in our data. For example, relative clauses some difficulties for the program, resulting in false identification of errors such as the following examples:

Ex3. *Students do not need to carry lots of books* ***which are*** *really heavy.*

Ex4. *They also use it for reading instead of printing* ***which is*** *easier and help them to read on more additional topics.*

Ex5. They can work on other things ***that give*** *more advantage on study.*

However, a future version of this program will incorporate Wh-pronouns and determiners as well as their dependency structures to account for these types of errors. Another error that appeared was related to the use of non-finite verbs, which were tagged as having subject-verb agreement errors.

Ex6. It would be waste of time for ***professors to teach*** *them how to use it*

Finally, some problems were simply a result of other, non-target, errors that are often present in learner language. In the following example the writer left out the pronoun "I" following *that*, which led the program to believe *that* is the subject of *have*.

Ex7. In the end, I believe ***that have*** *to be the one who decide what they want not the professors or schools."*

Overall, while there were a considerable number of false positives, the path forward for future improvement of the program's precision seems clear after identifying the sources of these problems. Recall, on the other hand, was relatively successful (though could be improved) and performed sometimes better than human raters, which is promising and points to the value of pursuing this project further.

References

Allen, M. (2017). Intercoder reliability techniques: Fleiss system. *The SAGE Encyclopedia of Communication Research Methods*. Thousand Oaks, CA: SAGE.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999).*Longman grammar of spoken and written English*. Harlow, England: Pearson Education.

Dulay, H., Burt, M., and Krashen, S. (1982). *Language two*. New York: Oxford University Press.

Ellis, R. and Barkhuizen, G. (2005). *Analyzing learner language*. Oxford: Oxford University Press.

Gass, S., Behney, J., & Plonsky, L. (2013).*Second language acquisition*. New York: Routledge.

Hayes, A., & Krippendor, K. (2007). Answering the call for standard reliability measures for coding data. *Communication Methods and Measures*, *1*(1), 77-89.

ISU EPT Corpus of Learner Writing (Release 2.2). (2018). Corpus compiled by the Applied Linguistics and Technology program and Bethany Gray at Iowa State University.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Qi, P., Zhang, Y., Bolton, J., and Manning, C. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Association for Computational Linguistics System Demonstrations*.

Stapa, S. H., & Izahar, M. M. (2010). Analysis of errors in subject-verb agreement among Malaysian ESL learners. *3L: Language, Linguistics, Literature*, *16*(1).

Singh, C. K. S., Singh, A. K. J., Razak, N. Q. A., & Ravinthar, T. (2017). Grammar Errors Made by ESL Tertiary Students in Writing. *English Language Teaching, 10*(5), 16-27.

SVA Tagger. (2020). Available at [1] https://github.com/Jamesetay1/520/blob/master/SVA/SVA.py and [2] https://github.com/Jamesetay1/520/tree/master/SVA

Swert, K. (2012). *Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha*. University of Amsterdam.

Wee, R., Sim, J., & Jusoff, K. (2010). Verb-form errors in EAP writing. *Educational Research and Reviews, 5*(1), 016-023.

Zheng, C., & Park, T. J. (2013). An analysis of errors in English writing made by Chinese and Korean university students. *Theory and practice in language studies, 3*(8), 1342.